1

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE

READ INSTRUCTIONS BEFORE COMPLETING FORM

1. REPORT NUMBER
AFIT/CI/NR 88- 115

2. GOVT ACCESSION NO.

3. RECIPIENT'S CATALOG NUMBER

4. TITLE (and Subtitle)
A SYSTEMS APPROACH TO THE AEROMEDICAL AIRCRAFT ROUTING PROBLEM USING A COMPUTER-BASED MODEL

5. TYPE OF REPORT & PERIOD COVERED
~~MS~~ THESIS

6. PERFORMING ORG. REPORT NUMBER

7. AUTHOR(s)
DENNIS ROBERT MCLAIN

8. CONTRACT OR GRANT NUMBER(s)

9. PERFORMING ORGANIZATION NAME AND ADDRESS
AFIT STUDENT AT: UNIVERSITY OF CALIFORNIA - BERKELEY

10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS

11. CONTROLLING OFFICE NAME AND ADDRESS

12. REPORT DATE
1988

13. NUMBER OF PAGES
427

14. MONITORING AGENCY NAME & ADDRESS(if different from Controlling Office)
AFIT/NR
Wright-Patterson AFB OH 45433-6583

15. SECURITY CLASS. (of this report)
UNCLASSIFIED

15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)
DISTRIBUTED UNLIMITED: APPROVED FOR PUBLIC RELEASE

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)
SAME AS REPORT

18. SUPPLEMENTARY NOTES
Approved for Public Release: IAW AFR 190-1
LYNN E. WOLAVER Lynn Wolaver 20 July 88
Dean for Research and Professional Development
Air Force Institute of Technology
Wright-Patterson AFB OH 45433-6583

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)
ATTACHED

88

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

TO WHOM IT MAY CONCERN 2 May 1988

As the copyright holder for the document "A Systems Approach to the Aeromedical Aircraft Problem using a Computer-based Model", I give my permission to the Military Airlift Command to reproduce this document for whatever official purposes the command wishes to use the copies for. Any of these copies provided to contractors should be given with the understanding that they are not authorized to reproduce them. Contractors may obtain additional copies from University Microfilms, Inc., Ann Arbor, MI.

Dennis R. McLain
Lt Col, USAF
Chief, Modernization Branch
Technical Support Division
JCS/J-8

[illegible]

A Systems Approach to the Aeromedical Aircraft Routing Problem using a Computer-based Model

By

Dennis Robert McLain

B.S. (United States Air Force Academy) 1968
M.S. (University of California, Los Angeles) 1969

DISSERTATION

Submitted in partial satisfaction of the requirements for the degree of

DOCTOR OF PHILOSOPHY

in

Business Administration

in the

GRADUATE DIVISION

OF THE

UNIVERSITY OF CALIFORNIA, BERKELEY

Approved: ....C. West Churchman.....7/10/84....
Chairman Date

....Arie Segev.....7/5/84....

....C. Roger Glassey 8/1/84.....

....

A Systems Approach to the Aeromedical Aircraft Routing Problem using a Computer-based Model

by

Dennis Robert McLain

DISCLAIMER

This dissertation represents the views of the author and does not necessarily reflect the official views of the Military Airlift Command, the Department of the Air Force, or the Department of Defense.

TO SANDY, SARAH, AND THE BEEJ

ACKNOWLEDGEMENTS

It is very difficult to find words that adequately thank all those people who have been so helpful and supportive. More than anyone else, my wife Sandra has never let me quit, and has assumed an unfair share of too many burdens to allow me the opportunity to complete my work. That the degree will be in my name only is one of life's major unfairnesses. I have gauged my progress by the growth of my daughters Sarah, born nine days before my first day in the program, and Stephanie, whose birth coincided with the beginning of the dissertation. My only regret in completing it is the time that it took at their expense.

Perhaps my most fortunate experience was to have had Professor West Churchman as my guide and mentor, for his extraordinary intellect and considerable patience and understanding have been primarily responsible for my completing the program. Through his teachings I have come to realize how very little we know about designing complex social systems, and how critically important it is to better understand and develop appropriate methods to improve them. His seminar and writings on systems philosophy have been a constant source of inspiration throughout my program. I offer special thanks to Professors Roger Glassey and Ari Segev for consenting to help when help was really needed.

I have also valued the friendship and assistance of many others, including Professors Austin Hoggatt, Ernest Koenigsberg, and my fellow students at the University of California upon whom I prevailed for far many more favors than I returned. Harry Psaraftis's work on the Dial-a-Ride problem inspired the many-to-many model. Larry Bodin and Bruce Golden gave unselfishly of their considerable knowledge and experience in vehicle routing modeling.

My program was initially sponsored by the US Air Force Institute of Technology, to whom I am grateful for scholarship support and, more importantly, for the opportunity to pursue the degree. I also thank those people at the Air Force Academy and at Scott Air Force Base who helped me overcome the difficulties of finishing my work. I also received first rate assistance from the Military Airlift Command Surgeon's staff and the 375th Aeromedical Airlift Wing, whom I hope will realize the benefits of this research. Finally, I offer special thanks to those who taught me the value of critical thinking and having the tenacity to stick with something worth doing and the pride to do it as well as I could. That would certainly include my parents, teachers, and all those I have been associated with in the Air Force over these past 20 years.

TABLE OF CONTENTS

TABLE OF CONTENTS (Continued)

Page

TABLE OF CONTENTS (Continued)

LIST OF TABLES

Table Page

LIST OF TABLES

LIST OF FIGURES

LIST OF FIGURES (Continued)

Figure Page

LIST OF FIGURES (Continued)

LIST OF FIGURES (Continued)

ABSTRACT

McLain, Dennis Robert. Ph.D., University of California, Berkeley, December 1984. A SYSTEMS APPROACH TO THE AEROMEDICAL AIRCRAFT ROUTING PROBLEM USING A COMPUTER-BASED MODEL. Chairman: C. West Churchman.

This research concerns transporting medical patients on specially-equipped aircraft from one medical facility to another, primarily to provide them treatment not available at the first facility. In addition to finding improved routing methods so that patients can be moved as directly and expeditiously as possible, we address other important issues by means of a systems approach, such as the limited numbers of aircraft and crews that restrict the ability of the system to provide direct or even same-day service.

The most extensive aeromedical system, operated by the US Department of Defense (DOD), evolved from World War II and Vietnam War casualty evacuation. It has two distinct and overlapping roles, preparing for wartime casualty movement, and the peacetime transportation of active duty military and other eligible clients. Of the world-wide DOD network, we examined only the continental US portion, and we were primarily concerned with its peacetime operation.

The purpose of the thesis is to design an aeromedical planning system that will schedule weekly regional service, and produce daily routings. Historical patient movement data provides information that can be used in regional service planning. At the root of routing and sequencing is

the combinatorially difficult problem of finding solutions that satisfy the ordering restriction that patients be picked up before they are delivered. Routing problems which require both pickup and delivery service are commonly called many-to-many, two-ended service problems. We present solution methods for both multiple depot and multiple aircraft many-to-many problems.

Aeromedical service is demand-responsive; the decision to move a patient and the choice of a destination are made a priori and independent of the flight planning process. Both advance (or subscriber) and immediate (dynamic) service requests are generated. The DOD system is similar in many respects to urban paratransit systems for the elderly and handicapped, called 'Dial-a-Ride' systems, but differs in that patients are moved according to their medical condition. Patient deliveries are not made to meet scheduled appointments, so there are no service time constraints. Other differences stem from the use of airplanes instead of ground vehicles.

In developing a series of planning models, we encountered increasingly difficult problems, particularly in incorporating the many linkages that exist among organizational subsystems. Future research is needed to advance the theory of mathematical programming to permit the development of models that adequately represent the entire system, an elusive but worthwhile ideal to seek.

CHAPTER I

INTRODUCTION

Although modern transportation systems have become increasingly complex and technologically sophisticated, their fundamental function remains the same: to move something from one location to another. We move ourselves and our goods and services, economists say, because what we want is not always where we want it when we want it. When transportation increases the capacity of goods and services to satisfy human wants, by resolving these problems of time and place, they say that our utility has increased.[1] When we improve our transportation systems so that we are able to both increase utility and decrease the adverse effects of doing so, we say that we have progressed.

The principal focus of this thesis is on a medical transportation system that uses specially equipped and modified aircraft to resolve certain time-and-place disutilities that arise in the treatment of patients by the US Department of Defense (DOD) Military Health Services System. With hospitals and clinics located at many, widely dispersed DOD installations and eight million potential beneficiaries, the economics of health care delivery, shortages of medical specialists and equipment, and the need for extensive training programs for wartime preparation have forced DOD to consolidate and specialize its facilities. As a result, DOD cannot deliver adequate health service in all specialties at every installation.

Air transportation greatly alleviates the resulting health service distribution problems. The time utility of medical treatment is obviously greatest at the time an injury or illness occurs, and declines as treatment is delayed. Air transportation greatly reduces the time, over other modes, to transport patients to medical facilities for urgently needed care. In some instances, the distance or terrain over which patients must be moved precludes travel on any other type of vehicle. Alternatively, aircraft can rapidly move treatment teams and their equipment to victims too critically injured or ill to be moved. In terms of place utility, air transportation is unquestionably important in linking DOD facilities in 35 countries, and in permitting the specialization and aglomeration of facilities that provide significant economies of scale, achieve desired staff training goals, and allow the elimination of uneconomical facilities.

The DOD patient transportation problem is considerably more complex than finding the fastest means to move patients from one point to another. Given the distribution of treatment resources and individual patient treatment needs, each patient should be transported to his destination medical facility as expeditiously as possible. However, given a limited number of aircraft, flight crews, and specially trained medical teams, a mass transportation approach, rather than air taxi service for each individual patient, must be used in order to move 100 or more patients each day. Unlike many mass transportation systems, each individual patient orgin and destination must be served, which precludes using a fixed route network. Despite the difficult plannng problems that result, the system moves more than 70,000 patients each year.

But, the problems associated with operating modern jet aircraft, particularly the escalating factor costs of fuel, flight and medical crew salaries and training, and maintenance manpower and materiel, continually challenge system managers to maintain positive utility differentials. Despite their high initial costs, new aircraft that fly faster, improve comfort, carry more patients, and reduce fuel consumption can greatly improve patient service and reduce operating costs. For example, McDonnell Douglas, manufacturer of the DC-9 aircraft currently used by the DOD medical air transportation system, claims that its latest DC-9 Super 80 model is considerably quieter, flies greater distances, and uses 30 per cent less fuel than the model now in use. [MCDO84] Alternatively, 'retrofitting' existing aircraft with more fuel-efficient engines, electronic fuel management equipment, and structural modifications can achieve comparable improvements with smaller capital outlays. Increasing fuel-carrying capacity, reducing fuel consumption, and improving aircraft reliability can reduce or eliminate ground servicing, maintenance delays between flights and aircraft out-of-commission time. Besides improving service to individual patients, these changes can increase the system's capability, by effectively increasing aircraft fleet capacity without changing fleet size. Technological solutions such as these, then, are indisputably important in achieving progress.

But it is entirely possible, as Schon contends in Technology and Change [SCHO67], that we have come to expect technology, almost as a matter of blind faith, to provide the means to progress, to the point that organizations such as DOD routinely devise plans for new aircraft or improvements to existing designs that call for materials, construction methods, and components that we haven't even invented yet. That faith has not always been upheld. But even when it has, successful technological solutions may blind us to other problems and solutions because of one overriding reason: we fail to consider the whole system, and focus on just its engineering or technological aspects.

In this study, we examined the DOD medical transportation problem from a systems perspective. As our focus shifted from improving transportation technology as the sole means for achieving progress, to such concerns as the nature, causes, and structure of patient movement, and the functioning of the larger health care delivery system within which this transportation service operates, we identified a number of significant, non-technological problems. For example, we observed that because of the their high costs, DOD can afford few aircraft relative to the number of patients to be served. Because of this, the most basic medical air transportation problem is to determine the order in which each aircraft stops to pick up and deliver patients. On further examination, we found that this difficult combinatorial problem, if not solved adequately, can easily negate the value of using even the most technologically efficient aircraft.

As we expanded our framework from single aircraft routing problems with few individual patient movements, to moving a significant number of patients with a fleet of aircraft, we discovered additional problems. Multiple aircraft create the so-called partitioning problem, of deciding to which aircraft to assign patients. And, we discovered that chosing where each aircraft ended its route significantly effected system operation the following day. So too did selecting which regions of the United States to serve on a given day, and which ones not to serve.

Further examination revealed that, when the spatial separation of patient locations was coupled with the limited number of aircraft and daily operating hours, we could not always move every patient from his origin directly (or even indirectly, on the same day) to his desired destination. This required us to develop effective service criteria,[2] rules for 'storing' patients overnight at enroute locations,[3] and schemes for determining not just the sequence of stops on one flight, but interrelated sequences on several flights to final patient destinations. Otherwise, patients experienced excessive times waiting at their origins and made unnecessary enroute stops.

Looking beyond the aeromedical subsystem to the larger health care delivery system within which it operates, we found that these problems are in turn compounded by institutional idiosyncrasies, structural characteristics of patient movement, and other factors. For example, we found that rules and operating policies bias the selection of patient destinations; facilities operated by the same military branch are favored over others geographically closer. Patient service tends to be demand-responsive, owing to the unpredictable nature of medical needs and to the policy of not rigidly scheduling aircraft routes. The relative concentration of medical services at a few large installations, and the myriad origins that stem from a number of uncoordinated health benefit programs, produced a many-to-fewer, origin-to-destination movement structure.

By using a systems approach, then, we attempted to treat the aeromedical transportation problem as comprehensively as possible, rather than simply isolating one or two problem areas. We found, as others have from the time of Anaxagoras, that a complex system such as this is not decomposable into less complex subsystems for two principal reasons. First, the subsystem problems are difficult in their own right. And more significantly, linkages between subproblems directly determine whole system performance, and cannot be ignored, nor easily handled. Our system model attempts to explicitly include linkages between subsystems. Admittedly, in the end we only solved a few technical aeromedical management problems; we hope that this investigation provides a framework and starting point for others to attack the problems we did not solve.

1.3 **Problems Addressed**. In this study our principal concern is the aeromedical transportation problem, which at its highest level of abstraction is to:

> Given a set of n patients located at origin hospitals, and a fleet of m aircraft, determine the 'best' set of routes that will transport them to destination medical facilities, subject to environmental and operating constraints.

We first developed a conceptual model of the DOD medical air transportation system. Perhaps the most difficult problem in doing so was to (1) identify the real objectives of the system, and (2) choose adequate measures of performance in achieving those goals. We found the system justified largely in terms of one objective, preparation for wartime operation, and its performance measured primarily in terms of a health benefit goal.[4] These problems were exacerbated by two traits common to public sector organizations:

insulation from market forces, and a lack of private sector motives such as cost minimization or profit maximization.

With an understanding of the purpose and functioning of the system, we then addressed the nature of patient movement. Specifically, we used various analytical methods to determine movement structure and patterns. The hierarchical, regional structure, and two-tiered flow patterns that we found are the main foundations of our system model.

The third problem we addressed is the design of an aeromedical transportation planning subsystem. Churchman notes that a planning subsystem

> has to deal with the generation of plans for the system. ... The management sets the component goals, allocates the resources,and controls system performance. Not only does the management of a system generate the plans of a system, but it also must insure that the plans are being carried out in accordance with its original ideas. This activity is often called "control". However, control does not only mean the examination of whether plans are being carried out correctly; it also implies an evaluation of plans and consequently, a change of plans. [CHUR69]

Aeromedical transportation plans take the form of routes to pick up and deliver patients, and weekly and monthly schedules specifying which regions of the country will be served each day. The planning problem involves three related and difficult questions, which we investigated.

The <u>optimal routing problem</u> is to find the best set of daily regional and interregional routes to service patient origins, destinations, and overnight stopover points called aeromedical staging facilities. The two critical questions to be answered are (i) which nodes will be visited, and (ii) in what order should they be visited. Routing problems with origin-to-destination service requirements are common in practice, but adequate solution methods are not, which forced us to find new techniques.

In the routing design problem, we must determine for several operating periods which major stops to visit on which days, without regard to visit order. Optimal routing and routing design are closely related; daily routes are particularly sensitive to the first and last stops of each aircraft, which the routing design determines.

According to Churchman, aeromedical transportation managers should, as part of their control activity, see to the

> construction of management information systems (MIS) that will record the relevant information for decision-making purposes and specifically will tell the richest story about the use of resources, including lost opportunities. [CHUR69]

To design such an MIS we followed the premise that the data it needs to record is specified by the routing and routing design decision models it needs to support. We investigated ways to integrate data and routing decisions into a resource allocation model. The model we derived utilizes a method of coordinating the allocation of resources (principally aircraft and crews) to improve client service.

1.2 **Major Contributions.** The model we developed makes two contributions to transportation management theory:

> (1) We will show that a theoretical model of resource allocation can be used to suggest daily schedules and changes to them, and to indicate improvements in longer-term resource use. Because rational models such as these are sensitive to improper parameter values and misspecifications of or changes in problem characteristics, we will demonstrate that control information derived from experience can be used to improve system performance. This enables the rational model to adapt to a dynamic environment, which its static formulation otherwise could not.
>
> (2) As a major contribution to vehicle routing theory, we present new branch-and-bound algorithms that solve for optimal routing. The routing problems cannot be solved directly by integer linear programming (ILP) methods because of the limitations of available ILP methods.

1.3 **Major Purpose of the Thesis.** The major purpose of the thesis, then, is to contribute to the understanding of resource-allocation decision making by addressing two aspects of the aeromedical transportation problem: how we should devise an MIS that will provide information and even suggested planning decisions to aeromedical transportation managers, and how to construct its underlying analytical model. Our major thesis is that the information flows and decision making of the resource allocation process can be modeled as a linear program, and solved by a resource-directive coordination method derived from the theory of decomposition in mathematical programming. While an extensive theoretical literature discusses models of this kind, applications such as ours are rare.

The aeromedical transportation organization possesses a highly formalized resource allocation process, and a relatively easily quantified set of patient movement requirements, resources, constraints, and standard operating procedures that are compatible with the severe assumptions of mathematical programming models. And, the resource-directive decomposition method appears to be, among various kinds, most analogous to observed organizational behavior. From an implementation standpoint, our choices reflect our agreement with Atkins that

> [. . .]it is not our purpose to claim that one [decomposition] method is better than others, but rather that some are more suitable depending on the circumstances. Our intention is that if we have any particular organization in mind and if we can identify key parameters that reflect the style of management in that organization, and also if we believe that we are unlikely to have success implementing or getting new planning tools accepted that cut across existing managerial styles and prerogatives, then we can sharply narrow down the choice of decomposition procedures that we might be tempted to use as analogies. [ATKI74]

1.4 Organization of the Thesis. Chapter II describes the aeromedical transportation system and the planning problem to be solved, based upon an examination of the organization and actual system operation over a three-month period. Chapter III addresses the design of rational models of organizational resource allocation, presents the theoretical concepts of resource-directive decomposition, reviews the relevant literature, and presents a model of the aeromedical planning problem. Chapters IV and V discuss aircraft routing problems and new solution algorithms. Chapter VI introduces the full allocation model. Chapter VII summarizes our findings and suggests future research.

ENDNOTES

1. Economists ascribe the increase to changes in time utility, how closely a commodity (or service) is made available relative to its time of greatest usefulness, and place utility, how closely a commodity is located relative to its place of greatest usefulness. [LOWE75]

2. In planning to pick up, transport, and deliver patients, we must decide if we will provide complete, partial (to or from an intermediate point), or no service. Criteria could be geographical (serving only selected places), medical (by diagnostic category or urgency), ethical (based upon notions of equitable treatment), institutional (observing policies and regulations), or facilitational (moving a patient toward, not away from, his eventual destination).

3. These rules include stipulations such as which diagnostic categories and movement precedences permit or prohibit overnight stops, and the maximum number of stops.

4. This anomaly occurs in civil defense and natural disaster medical planning problems. Justifying a system large enough to provide adequate services under emergency conditions is extremely difficult if the economic costs to create and sustain excess capacity greatly exceed those incurred in normal circumstances.

costly[2] for general use, but in severe trauma cases, they significantly reduced the fatality rate. [HELI83] The experience of Maryland Medivacs, operated by the Maryland State Police, bears this out. "Since the program began in Maryland [in 1970], 85 per cent of victims with life-threatening injuries medivacced [sic] to this [University of Maryland's] trauma center lived." [CBS84] However, they

> "are not inherently better than ground transport vehicles for patient care in a given emergency medical system. Their value over ground transport increases proportionally to the time saved in transport from site A to site B and the level of care enroute." [HELI83]

In emergency and non-emergency situations, the value of aeromedical transportation systems he refers to stems from:

1) Supporting and taking advantage of larger, more specialized and regionally centralized medical facilities and staffs. [JOHN77]

2) Providing an economic alternative to maintaining fixed facilities in sparsely populated areas.

3) Greatly expanding the coverage of emergency medical services, literally representing the difference between life and death for those who are gravely injured or become severely ill, particularly in remote areas.

4) Directly exploiting aircraft speed and independence of the underlying terrain. Fixed wing aeromedical aircraft are roughly five times as rapid as the fastest train and eight times as speedy as the fastest vessel. [JOHN76] While slower, helicopters are still two to three times faster than ground transportation.

5) Providing further benefit to the patient through the elimination of the discomforts and stresses inherent in surface transportation and the frequent stops and the accompanying bumping and jolting. [JOHN77] As one authority says,

CHAPTER II

THE AEROMEDICAL TRANSPORTATION SYSTEM DESIGN PROBLEM

Our destination is never a place but rather a new way of looking at things. Henry Miller

2.1 Introduction. Aeromedical transportation[1] is not new. More than 100 years ago, during the siege of Paris in 1870, casualties were first evacuated by air, using observation balloons. World War II saw the first major aerial evacuation efforts, when converted cargo planes moved casualties from field hospitals and aid stations to better-equipped facilities. By the end of the war, the military routinely flew patients to the United States from the Pacific and European Theaters. During the Vietnam War, fewer than one per cent of those casualties who were evacuated to a medical facility died. [DEPA78a,p.4-7] Using helicopters and transports converted to carry patients, the US military medical system could move patients, depending on their needs and condition, from the battlefield to the most suitable treatment facility, including ones in the US.

Until recently most systems were designed to support military operations, but in the past eleven years, some 55 non-military, hospital-based helicopter ambulances have been used in more than 24 states in the US to apply battlefield casualty evacuation equipment and methods to the emergency movement of severe trauma cases. [HELI83,REIC82] As one study recently revealed, such vehicles are too

"Patients cannot be subjected to the use of vehicles for transportation that leave them uncomfortable, apprehensive, or exhausted." [MCFA53]

6) "Reaching patients who either could not be reached at all by ground transportation or would have to wait too long." [HELI83] "Almost any patient who can be transported at all, can be moved by an aircraft which is suitably equipped and has medical personnel aboard knowledgeable about physiological changes patients may experience in flight." [JOHN77,p.452]

7) Alternatively, when serious injury or illness prevents movement, getting a higher level of medical expertise direcly to the site. [HELI83]

8) With respect to routing and frequency of operation, very flexibly adapting to the requirements of the patients.

Aeromedical transportation systems differ widely in purpose, in the types of services they provide, and in the number of cases they handle. This study examines the US Department of Defense (DOD) aeromedical transportation system, the largest, most experienced, and least specialized, in terms of the area it serves and the number and types of cases it handles. Private and non-military public sector systems are considerably smaller. Air-Evac International, for example, serves only the cities of San Diego and Houston. Maryland Medivacs covers the state of Maryland. Flight for Life, the emergency aeromedical transportation branch of St. Anthony's Hospital in Denver, specializes in transporting critically injured and acutely ill to and linking remote areas of Colorado, Wyoming, Montana, North and South Dakota, and Nebraska with, large metropolitan Denver hospitals. The DOD system moved an average of 5000 patients per month in 1981 in the US alone [JONE82], while Air-Evac handled an average of only 14 to 16 patients per month in 1981. [SCHI81] Maryland Medivacs transported 2530, primarily accident victims. [CBS 84] All 42 hospital-based systems in the US combined handled 30000 patients in 1981, half the number moved by the DOD system. [REIC82,JONE82]

Besides providing both emergency and non-emergency patient transportation, the DOD system is also interesting from the standpoint that it is designed to accomplish two very diferent missions, in peacetime and in time of war. The peacetime system operates a worldwide network of aircraft routes and patient handling facilities. The wartime system will use the same resources and methods, but will be expanded with wartime mobilization of reserve flight and medical personnel. Current wartime planning calls for helicopters and surface vehicles to move casualties from the battlefield to aid stations and theater hospitals. For more serious cases, C-141B transport planes[3] are to be used to carry patients from a war theater to the US. C-9A's would then deliver patients to their destination hospitals if arrival bases did not have adequate bed capacity or medical service capability.

The DOD system posed some particularly difficult challenges to our study. First, because it must satisfy two interrelated objectives that both conflict with and compliment each other, two decidedly different planning

2.1.1 Military Health Care. Public law establishes two major objectives for the DOD Military Health Services System. [DRMS79] The first, which we will refer to as the readiness objective, is to maintain the health of over two million active duty US military personnel, to ensure they are physically fit to perform wartime duties, and train a medical corps to treat the sick and wounded in wartime. The second provides health benefits to those eligible under federal statutes. In Fiscal Year 1977,

> "2.1 million active duty military personnel, 2.9 million dependents of active duty personnel, 1.2 million retirees, 2.3 million dependents of retirees and 0.4 million survivors of members or former members" were eligible for free or low-cost treatment by the US military health care system."[4,5] [DRMS79,p.80ff]

We refer to this as the benefit objective.

The two objectives are neither in complete conflict nor are they totally separable. Since the establishment of DOD in 1947, family-oriented, non-war-related health benefit services have constituted a substantial proportion of all care provided. During the same time period, at the same hospi-tals and clinics, the same medical staffs providing benefit-oriented care have also maintained the peacetime health of active duty soldiers. Thus, the same means serve both ends, and we cannot separate objective achievement into two parts. And, treating the active duty client group achieves both objectives simultaneously.

problems must be solved. While we focused on peacetime planning, the linkages with wartime planning were substantial. Secondly, from a systems perspective, DOD aeromedical transportation is only part of, and is inextricably tied to, a larger military health care delivery system, which in turn is only part of one federal governmental subsystem, and so on. To understand the DOD aeromedical transportation system, we could not ignore the larger systems within which it operates. At the same time, however, comprehensiveness implies endless expansion of the system's boundaries, and little hope for analysis.

To resolve this dilemma we argue that the most appropriate approach is to use judgement to set some tractable boundaries, use observations of systems performance to derive a conceptual model of how the system ought to work, and develop a method that has the potential to improve the service the system should provide. Our goal is two-fold: to make system improvements (although we don't know how long that will take or to what extent we will succeed) and to gain experience and learning from attempting to understand how the system works. We first describe the US Military Health Service System, and how the need for aeromedical transportation arises. We then use a systems approach to learn how the current system works, using actual patient movement data to analyze system demands, and conclude with a statement of the system design problem we will address.

Non-separability of the objectives arises in other ways. First, the benefit objective tends to act as a compensation device. That is, free or very low-cost medical care can be perceived by the soldier as form of in-kind compensation. This perception can be a significant factor in the soldier's decision to continue his military career, which contributes indirectly to the readiness objective through the retention of a trained soldier.

The strong interrelationship between the two objectives is also evident if one considers the most demanding health care requirement, treating sick and wounded soldiers in wartime. A principal function of the wartime system is to transport those wounded soldiers whose expected treatment and convalescence time would exceed a specified limit to US hospitals. Daily peacetime operation of the aeromedical system contributes directly to the benefit objective by transporting those eligible to receive care between treatment facilities. At the same time, both the operators of the aeromedical system and medical staff at the sending and receiving hospitals rehearse their wartime patient movement roles, test new equipment and procedures, and continually exercise the command and control system.

Conflict between the objectives also exists, particularly in training military health care professionals, and in organizing and positioning military medical services to achieve both objectives simultaneously:

> Tailoring the health care system for two missions, a peacetime one and a wartime one, poses difficult problems. An ideal wartime system would consist of a physician force heavy in surgical skills, well prepared to deal with trauma, and a number of large hospitals in the United States concentrated near evacuation points.[6] An ideal peacetime system would consist of a physician force heavier in pediatricians and other primary care physicians located in smaller facilities at each military installation. [DRMS79,p.80]

Indeed, the two objectives pose two fundamentally different planning problems. The peacetime problem, given a relatively (1) fixed health care delivery system in terms of staffing and location of facilities, (2) stable demand for medical services, (3) stable budget, and (4), perhaps most importantly, considerable past experience, is to establish a health services system that satisfies both objectives and fulfills identified training needs.[7] Despite reported shortages of military physicians, there is no conclusive evidence of either a positive or negative impact on health care delivery. [DRMS79]

On the other hand, the problem of designing a wartime system, without adequate estimates of (1) casualties; (2) facilities and staff requirements; (3) transportation needs both within the theater and to the United States; and (4), CONUS hospital bed availability;[8] one that is, to some extent, economically feasible in peacetime, and yet is able to quickly adapt to dramatically increased requirements at the outset of a conflict, is considerably more complex. The complexity of the design task is evident when we consider the various functions the system must perform to achieve that goal.

The most important wartime role of the system is to preserve the largest possible supply of manpower. There are two primary ways to accomplish this. When patients do not require extensive and complex procedures, long convalescences, or sophisticated equipment not found in battlefield aid stations and hospitals, treating them within the theater returns them as quickly as possible to duty. The alternative is to defer treatment until patients can be evacuated to the United States. (In addition, the wartime military health care system must continue to provide health care to eligible non-combatants, e.g., family members of active duty soldiers, by shifting responsibility to the civilian sector, a problem the DOD has only recently addressed. [DEPA81])

These two alternatives create what the Defense Resource Management Study called the evacuation planning problem:

> The most important unknown is the number and type of sick and wounded. Given patients, the next problem is to determine what kinds of medical resources will be required to care for them in an acceptable way. Given the kinds of resources needed, the next question is where they should or can be obtained. Here evacuation policy is crucial, because it is the variable which determines what facilities have to be located in the theater. This, in turn, determines options for using CONUS military, federal, and civilian hospitals and personnel. ... Evacuation policy affects virtually every aspect of contingency planning from airlift to engineer construction requirements.[DRMS79,p.85]

An evacuation policy stipulates that, if a patient cannot be treated and returned to duty within a certain number of days, he will be evacuated. Establishing an evacuation policy is very difficult. Independent variables such as the type of warfare (static vs. fluid force movements, weaponry used, terrain, etc.), availability of patient transportation, and the quantity and distribution of medical resources constrain the set of feasible policies. An evacuation policy is also an independent variable, since it in effect determines the relative amounts of treatment in the theater, and evacuation for treatment in the U. S..

Treatment in a theater will be difficult because of severe shortages in physicians and facilities that are expected to exist at the outset of a conflict. [DRMS79] Shortages are to be relieved by deploying medical staff and equipment from the US to reinforce existing medical units and establish new units. But that solution requires estimates of casualties that are very difficult to make, in order to decide when, where, and to what extent additional resources are needed. And, shifting resources from the US requires facility administrators in the US to plan for operation without staff and equipment lost to deployment. The resulting loss of treatment capability may be so extensive that non-active duty patients would have to be transferred to other facilities, and arrangements for civilian hospitals to admit new patients may be necessary. [DEPA81]

In World War I, the evacuation policy was typically 150 days. By World War II, this time had decreased to 60 days, and to 15 days during the Vietnam War. [DEPA78] The Defense Resource Management Study points out that in

> 1974 ..., the Secretary [of Defense] directed the services and the JCS [Joint Chiefs of Staff] to plan to evacuate a larger proportion of wartime patients to CONUS then they had been planning to evacuate. To do this they were to shorten the evacuation policy, ... It was clear at that time, and it remains true today, that the services had programmed too little hospital capability overseas to treat the number of casualties expected to need in-theater treatment under the approved evacuation policy. Shortening the evacuation policy and thereby returning a greater proportion of patients to CONUS offered one solution. [DRMS79,pp.85-86]

Returning patients to the United States provides the alternative means to offset in-theater treatment shortages, but not without creating significant problems. Because the US cannot maintain large military reserves of men and equipment throughout the world for all possible contingencies, current strategy calls for deploying of troops and cargo from the US by air to a conflict. On their return flights to the US, transport aircraft would be converted to carry patients to US hospitals. (Currently, no other means, such as hospital ships, are available). But, using transport aircraft for aeromedical transportation creates at least three major problems:

1) As the US increasingly employs weapons with 'black boxes' that must be repaired in the US, these high priority items compete with patients for space on returning aircraft;

2) Cargo movements dictate aircraft routes and schedules. As a result, patients may arrive at US bases that have inadequate medical staff or facilities, or the distribution of patient arrivals and the availability of medical resources may be poorly matched over time.

3) Aeromedical equipment for converting transports to carry patients must be returned to the theater, competing with critically needed war materiel for space. (Of the 13 pallets of cargo a C-141B can carry, two or three would be required for aeromedical equipment). Using an alternative aircraft configuration with very little additional equipment alleviates the problem of recycling equipment, but substantially reduces carrying capacity and patient comfort.

The second problem is particularly germane to our analysis. A significant amount of patient redistribution within the CONUS may be required.[9] This is a major justification for peacetime operation of the CONUS aeromedical system, to prepare for large-scale patient movements to resolve care distribution and availability problems.

Differences between peacetime and wartime operation are clearly evident in individual patient care. In peacetime, primary emphasis is placed on individual patient needs. If a transfer is necessary, the basic policy is to send the patient to the nearest facility with the required capability. [SIVE78] However, several exceptions are permitted, including transfers to hospitals near patients' duty stations, places of residence or to the same facility where care had previously been given for follow-up examinations. And, transfers may be made to insure sufficient caseloads for training programs and for humanitarian reasons.

During peacetime, aeromedical transportation planners attempt to limit delays in initially picking up and in delivering patients, total in-transit times, and the number of overnight stops. Aircraft are dispatched to the airports nearest the patient's origin and destination; it is not uncommon for a single patient to be enplaned or deplaned at a given stop. Because of this, individual flight segments and even entire missions[10] may carry only one or two patients. Maximum efficiency goals, such as minimizing the total distance all patients travel, operating cost, or total aircraft flight time, conflict with service oriented toward individual patients.

In wartime, emphasis will be placed upon moving full patient loads as expeditiously as possible from evacuation points to the nearest facility with enough capacity and treatment capability. Patient volumes dictate this; in the first 30 to 60 days of a major conflict, more patients than the average peacetime annual total of 65000 [JONE82] may require transportation. Each C-141B transport bringing patients to the CONUS has a patient capacity two and one-half to four times that of the C-9A used for CONUS redistribution from evacuation points to destination hospitals.[11] With only eleven C-9A's available,[12] just four C-141B arrivals could conceivably require the entire C-9 fleet. Although each C-9 may be capable of making several redistribution sorties per day, an influx of several thousand casualties per day could severely overtax the C-9 subsystem. Under these circumstances, maximum redistribution flows would govern wartime movements.

In attempting to accomplish both its peacetime and wartime missions, the health care system is not ideally suited for either. Maintaining large hospitals near evacuation points implies allocating physicians away from primary care facilities at each military installation. Training programs at the large facilities require sufficient numbers of cases to sustain them. Both of these circumstances create a medical transportation requirement. And, the differences between peacetime and wartime patient movement, primarily in the emphasis on individual versus mass transfers, suggest that the aeromedical transportation planning function should be capable of supporting both missions. In the next section, we will discuss a number of additional factors that increase the need for aeromedical transportation, and complicate planning system design problems.

2.1.2 Military Health Services System. This section briefly describes the organization and facilities of the DOD Military Health Services System. We briefly discuss the nature of the health service delivery problem and hypothesize organizational institutional, geographical and economic factors[13] that create medical transportation needs. Readers interested in more detailed descriptions of the DOD health services system should see [DEPA83c] and [DEPA83d].

2.1.2.1 Organization. The decentralized organization of medical services in DOD mirrors the structure of DOD itself. DOD and its major components, the three military services (Army, Navy, and Air Force), utilize a line-and-staff structure.[14] Within each service, responsibility for medical matters is vested at all levels of the organization in medical officers who have de facto authority to administer the medical health program. Each service's Surgeon General manages the medical program of his service, to include the administrative management of all medical personnel, and the operation of facilities.

One apparent outgrowth of this structure is that Navy clinical staffs serve clients at Navy hospitals who are predominantly Navy personnel. In calendar years 1981, 1982 and 1983, the highest level of interservice aeromedical transfers was only 24.2 per cent. [HARR84] While no restrictions prevent members of one service from obtaining care at another service's facility, our patient movement analysis in a later section reveals a bias toward moving patients to facilities operated by the same service.

In addition to their administrative assignments, military health care facilities belong to geographical regions, presumably among other things, to limit the distances involved in patient transfers. However, no central authority exists within each region to coordinate planning and resource allocation, to include patient transportation. The Armed Services Medical Regulating Office (ASMRO), which centrally determines ("regulates") routine in-patient transfers between facilities, and validates requests for higher than routine movement precedence, is not required to observe regional boundaries. [SIVE78] Regardless, the Government Accounting Office has frequently criticized DOD for the number of interregional transfers ASMRO permits. [DEPA78b] We will examine patient transfers in detail later, and show that the structure of patient movements definitely exhibits a regional bias, even in the absence of official intraregional criteria.

Patient transfers depend significantly upon another, informal, organization among doctors. While ASMRO retains approval authority over routine in-patient movement requests, attempting to match patient needs with the closest available treatment, higher-than-routine precedence and out-patient transfers are handled by the hospitals and physicians concerned. That is, the administering physician who cannot provide needed treatment personally arranges care with a colleague at another facility. Presumably, a physician can veto arrangements he finds unsatisfactory, or suggest hospitals and physicians to ASMRO, which may explain the same-service bias to some extent. Thus, interpersonal relationships among physicians seem to intervene strongly in matching needed and available medical services.

Institutional rules also contribute to the same-service movement bias. Drug abusers and psychiatric patients must be treated by their own medical corps. Regardless of the nature of the care they require, patients who are also prisoners must be treated at facilities operated by their branch of service. [SIVE78]

Another important factor in military health services is the relationship between health care objectives and organizational structure. Responsibility for the readiness objective is clearly assigned to each of the military departments. DOD regulations stipulate that military members must be treated at military medical facilities.[15] On the other hand, other eligibles receive health benefits through a number of separately administered and uncoordinated programs. [DRMS79] Direct care delivery at a military facility is administered by the medical corps of the service managing the facility. Other medical treatments may be provided at civilian hospitals under the Civilian Health and Medical Program of the Uniformed Services (CHAMPUS), particularly when patients do not reside in a military hospital cachement area. CHAMPUS management responsibility is assigned to the Assistant Secretary of Defense for Health Affairs. In some circumstances, health care may by provided by the Veterans Administration (VA) or the US Public Health Service (USPHS), under negotiated reimbursement agreements.

Our patient movement analysis will show that these various treatment programs create many cases where an originating hospital transfers only one or two cases during the three-month observation period. Such non-routine cases complicate planning. Coordinating the patient transfer is more involved if the medical staff attending the patient at his origin is unfamiliar with aeromedical operating procedures. And, aircraft operations at seldom-used airfields involve a significant amount of preliminary coordination with airfield management, comparison of aircraft capabilities and limitations with airfield characteristics, and non-routine flight planning.

There is some evidence[16] to suggest that patient movements reflect a kind of organizational learning and adaptation. Doctors apparently time movement requests to coincide with aircraft movements, which do have a certain regularity, at least at the regional level. By timing a request, a doctor can increase the likelihood that his patient is picked up and delivered with minimum delay. This may be particularly significant in making scheduling alterations that destroy perceived regularities. Also, request timing closely mirrors the Monday-through-Friday hospital work week.

To this point, we have asserted that aeromedical transfers are made in response to requirements that can be influenced by:

1) Institutional arrangements that ensure patients are moved to facilities belonging to the same service;

2) Informal arrangements among doctors that strongly influence the choice of a destination facility;

3) Organizational learning that affects transfer request timing and choice of a destination facility;

4) Individual patient considerations, such as transfers to locations near duty stations and home towns;

5) Regional tendency toward nearby facilities or major regional centers and away from obvious flight scheduling difficulties;

6) Separate and uncoordinated care programs generating movement requirements from a large number of military, civilian, VA, and USPHS hospitals.

7) Medical training program requirements.

If valid, these factors suggest that predicting patient movements using simple correlations among origins, destinations, and diagnoses, and using them to predict patient transfers would be very difficult (and misguided). The next section will show that a number of other geographic and economic factors further complicate the DOD aeromedical transportation planning problem.

2.1.2.2 **Facilities**. The magnitude of the military health services program is quite evident in the following:

> The DOD operates 170 hospitals, of which 129 are in [the] CONUS and 41 are overseas. Tho Army operates 50, the Navy 37, and the Air Force 83. The normal[17] bed capacity worldwide is 37,069. Operating[18] beds total 20,650, of which 17,636 are in the United States. In [the] CONUS, occupancy rates[19] range from 49 per cent to 97 per cent; the DOD average is 73 per cent. Fifteen Army hospitals, 32 Air Force hospitals, and one Navy hospital are located in remote or underserved geographic areas in [the] CONUS. In addition, there are 302 free-standing clinics and dispensaries and 19 drug and alcohol rehabilitation centers." [DRMS79,p.81]

These facilities are located in all 50 states, and in more than 35 countries. [DRMS79] Major problems in providing health care do not stem from an overall lack of capacity; rather, other factors contribute to supply and demand imbalances and spatial dispersion in the DOD system:

1) Many military installations and their medical facilities are geographically isolated within the United States. Alternative sources of care (e.g., civilian hospitals) may not be nearby.

2) The increased use of costly and complex new equipment and procedures, coupled with a shortage of specialists and a lack of patients who require those services, make it infeasible to provide a complete range of medical specialties at every DOD medical facility. This has led to hospital specialization.

3) Certain specialties, teaching and training functions, and research activities have been consolidated at a few very large installations. (While the stated goal is economies of scale, other explanations may be more accurate).

4) Overall manpower ceilings, specialty shortages, and budget constraints necessitate tradeoffs to meet both health care objectives. The need for pediatricians, for example, has to be balanced against the need for surgeons, with resulting shortages in both specialties throughout the system.

5) The locations and other characteristics such as the size of military installations and their medical facilities are determined by national defense requirements and by other political, legal and historical factors. The resulting distribution of facilities creates situations where large metropolitan areas, such as Los Angeles, have few military hospitals with limited capacity to serve a large number of eligible patients.

The overall result of these problems of geographic isolation, distributed shortages, limited budgets, and supply-demand imbalances is the need for an alternative to direct care delivery in all specialties at every military hospital and clinic. The DOD medical transportation system that both permits and resolves the problems created by agglomeration, specialization, and spatial dispersion must also be able to resolve the inherent differences between its peacetime and wartime roles. And, it must be able to do so within the particular organizational and institutional framework of the DOD that governs and affects patient transfers. The following section will examine aeromedical transportation from a systems perspective to develop a conceptual model of the system and its responses to its dual roles and these institutional and organizational idiosyncracies, and outline a design for DOD aeromedical transportation planning to improve those responses.

2.2 A Systems Approach to DOD Aeromedical Transportation. At the outset, we said that the principal focus of this thesis is on DOD's domestic aeromedical transportation service. We said in the first chapter that we were principally concerned with designing an aeromedical transportation management system, and that a management system is primarily responsible for the generation of plans. Logic would have it, then, that this thesis is about designing (or redesigning) DOD's aeromedical transportation planning system. It is.

Moreover, we said that we wanted to improve the system. Improvement implies a pragmatic concern with the future consequences of implementing the plans we generate, that the condition of those served by the system is improved, rather than worsened, as a direct result of the plans made by the planning system we design. If we seek improvement by means of scientific method, we will be using, to employ Churchman's label, a systems approach. [CHUR69]

Then what, it is fair to ask, constitutes a systems approach to the design of planning systems? Churchman wrestled with this question in perhaps the two most important recent works on systems philosophy, **The Design of Inquiring Systems** [CHUR71], and **The Systems Approach and its Enemies.** [CHUR79] The orthodox management science or operations research approach would identify the planning design problem, gather data, formulate a model, solve it, then recommend and implement actions based upon model results, in that order. But why is the rate of failure so high for this approach, which seems so seemingly straightforward? Churchman contends that our first concern should be implementation, considering how and if we can achieve improvement, what form any benefit will take, and how and to whom it will be distributed. By first considering the question of benefits, we should be able to determine the eventual design goal and decide at the outset if and how we can attain it. And, we should not expect the steps in the design effort, whatever their order, to occur once; instead, we should expect as we proceed to learn more about the system that will cause us to repeat previous steps.

To begin to address the question of improvement in planning, we first need to examine the concept of planning itself. By first examining the nature of plans and planning activity, we assert that the essence of planning is inquiry, which we define as finding knowledge that provides the ability to adjust the behavior of the planning system to changing circumstances. Designing an planning system implies, in turn, two major design tasks, creating an <u>information system</u> responsible for gathering data, converting data into information, and communicating planning information to a decision maker, and a <u>decision making system</u> that uses information to make and revise planning choices. Churchman uses the term inquiring system to refer to these combinations of information and decision systems that give the decision maker the ability to adjust his decisions to changing conditions, particularly in finding ways to improve the contribution of the system's resources.

If inquiring systems require an information system, it has become abundantly clear in our attempts to build information systems that they must be more than simple fact collectors, processors, and distributors. The most important characteristic of the information system element is that it must adequately incorporate linkages between system components, in order to obtain data useful for planning. But as we explore the question of how to design information systems, we discover that very difficult design problems must be resolved.

Plans are not simply the fortuitous collection of just the "right" data; the processing of data into information and the conversion of information into decisions can involve the solution of highly intractable problems. Inquiring systems, then, require a decision making component. Plans must not only be conceived in thought; they must be communicated to decision makers and implemented through action. And through a control process, managers must adjust their plans in response to perceived system performance. In other words, both the information and decision components are involved in much more than simply specifying plans. If our eventual design goal is to design the "optimal" planning system, it seems then that we must specify (1) how to acquire knowledge (by designing the most appropriate information system); and (2), how to translate knowledge into action (by selecting the most appropriate method of making and implementing planning decisions), so that something or someone served by the system (identified by some means as the most appropriate beneficiaries of the service, of course) realizes an improvement in that service (as determined by the most appropriate way to detect if and how service has improved).

Unfortunately, when we examine any potential candidate planning system, even before we model a specific planning problem, we discover that it is virtually impossible for the candidate system to satisfy this criteria for system design optimality. We find ourselves doomed at the outset if our goal is to develop the 'ideal' planning model. Alternatively, when we more closely examine the nature of design, we find that by attempting to characterize a system in terms of its basic components and their interrelationships, and then redesigning one or more subsystems toward the goal of whole system improvement, we learn something significant about the relationship between our design and system performance. That learning is the principal object of the systems approach, and of this investigation.

For a starting point, we need a suitable definition for the term plan. Webster's Dictionary suggests several:

> **plan**: ... *3a: a method of achieving something:* a way of carrying out a design: DEVICE ... b: a method of doing something: PROCEDURE, WAY ... c: a detailed and systematic formulation of a large-scale campaign or program of action ... d: a proposed undertaking or goal: AIM, INTENTION ... **syn** PLAN, DESIGN, PLOT SCHEME and PROJECT can mean, in common, a proposed method of doing something or achieving an end. PLAN implies mental formulation of a method or form ... DESIGN adds to PLAN the idea of intention in the disposition of individual parts, often suggesting definiteness of pattern, or a degree of harmony or order achieved. [1729-1730]

The synonymous relationship of design and plan, in which intention, action and achievement of ends are implied, is important to our concept of designing a planning system. Let us say, for a first approximation, that plans are:

a) <u>counterfactuals</u>, historically-derived theories about the future: "If we take actions x_1, x_2, ..., x_n, then Z will result." Counterfactuals cannot be scientifically tested, because we cannot observe or test for the occurrence of the result before the plans are carried out, particularly when we are not allowed to experiment with the system. Rather, any faith we have that these conjectures will work must be based on past experience. [CHUR77]

b) <u>decisions</u>, of which activities we will implement, (and which ones we will not, for which there may be some lost opportunity, and possibly an associated opportunity cost);

c) <u>co-producers</u> of future states, in the sense that, in concert with the uncontrollable elements of the system's environment, they partially determine what will occur.

d) <u>purposeful</u> *and* <u>implementable</u> <u>intentions</u>. Because planning addresses the consequences of future states, plans are deliberate attempts to achieve a particular state. If we only deliberately attempt what we think we can feasibly do, then our plans are that which we feel we can transform from intellectual conception into action;

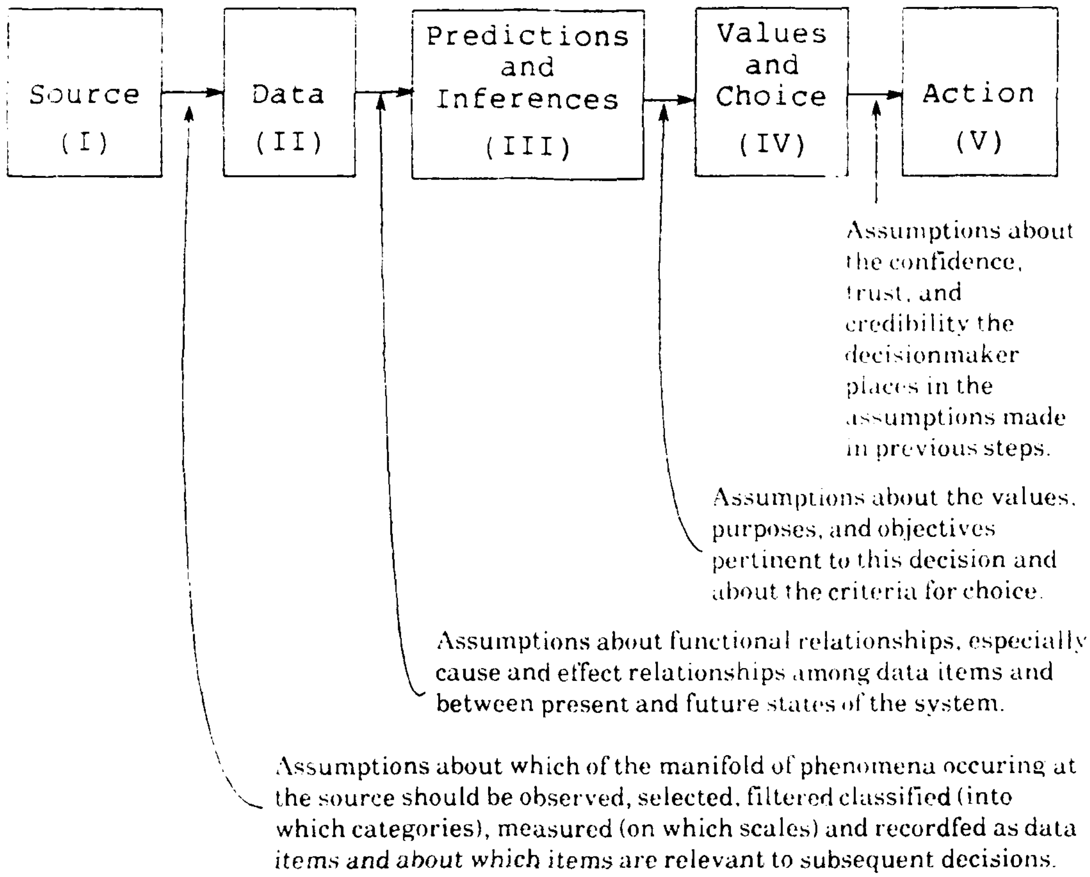
e) episodic and incremental. Changes in system demand, resource availability, and costs may require reformulation of plans. When aeromedical transportation plans are converted into actions, they co-produce results that directly influence future planning. Planning may have to correct past planning mistakes, and should anticipate the possibility of errors by specifying contingent alternatives.

f) products, of knowledge-based information systems. If data is any bit that describes something in the manifold of universal phenomena, and information is data used to construct and evaluate alternative future courses of action, then knowledge is information that allows the planner to distinguish between better and worse actions, in terms of changing system performance. In this sense, knowledge is used to produce planning choices.

To be able to generate plans, which must specify where every important entity in the system should be and when it should be there, every inquiring system requires a geometry (theory of space) and a kinematics (theory of time). The principal role of an inquirer's information system is to monitor the current system state, defined in terms of entity attributes, including location, at discrete instants in time. State data must be collected, filtered, stored, processed, transmitted, and ultimately, used to estimate or forecast future states via equations of system motion or transformation in time and space derived from the geometry and kinematics of the inquiring system. Its decision system, in turn, must decide on levels of controllable variables in the equations, carry out the resulting plans that are calculated, monitor and compare actual with planned outcomes, and generate any corrective actions necessary.

Mason's article, "Basic Concepts for Designing Management Information Systems" [MASO75], provides a particularly useful classification scheme (Figure 2.1) for planning systems, in which the classes depend upon the set of assumptions used to construct its information and decision making components. These assumptions range from those that prescribe what to observe, to those that characterize the motives and feelings of the decision maker, and they are the products of judgement's made by a system's management.

Assumptions present the first of many difficulties we will have in designing an optimal information system. Figure 2.1 suggests that the number of assumptions possible in an organizational setting is literally infinite. Churchman, Auerbach and Sadan, who identify four basic types of assumptions, descriptive, prescriptive, hypothetical, and categorical, show that even the number of all possible types is large, because these types are not exclusive. [CHUR75] Assumptions are not analytical, and they cannot be proven, because by their very nature, they are taken to be true. But, most significantly, they can be wrong. As Mason's scheme suggests, the outputs of one process stage are the inputs to the next. Therefore, since assumptions directly influence each stage, they ultimately influence



Assumptions about the confidence, trust, and credibility the decisionmaker places in the assumptions made in previous steps.

Assumptions about the values, purposes, and objectives pertinent to this decision and about the criteria for choice.

Assumptions about functional relationships, especially cause and effect relationships among data items and between present and future states of the system.

Assumptions about which of the manifold of phenomena occuring at the source should be observed, selected, filtered classified (into which categories), measured (on which scales) and recordfed as data items and about which items are relevant to subsequent decisions.

Figure 2.1. Mason's Taxonomy of Planning Systems. Source: [MASO75,p.13]

action. By deduction, erroneous assumptions can cause the wrong actions to be taken.

Mason's taxonomy implies that a wide variety of very different systems can be constructed, each one reflecting a particular world view, depending on the assumptions it embodies and the arrangement of its components:

Type of	Information System	Decision System
Planning System	Components	Components
Databank	I,II	III,IV,V
Predictive	I,II,III	IV,V
Decision-Making	I,II,III,IV	V
Decision-Taking	I,II,III,IV,V	

Databank systems, such as accounting systems and information utilities, are the simplest, most common, and provide the lowest level of decision making support. Predictive (e.g., econometric models) and decision making systems (e.g., optimization models) are relatively more complex, less common, and provide more sophisticated decision support. [MASO75] In a systems approach, an extremely important task is to determine which particular configuration constitutes the most appropriate inquiring system.

Linear programming (LP), a method we use in the thesis, provides an interesting example of a widely used, rational approach to planning. It provides us with a specific means by which to examine the question of appropriateness and some reflections on the phenomenological nature of planning. In Turban's survey [TURB72] of the corporate use of operations research models, LP ranked third after statistical analysis and simulation in frequency of use. Ledbetter and Cox [LEDB77] found in their study of the use of analytical models by 176 of Fortune's 500 largest corporations that LP ranked second after regression analysis. LP applications range from the simplest textbook problems to the central management of entire economies. [KORN67]

In the instruction of students of operations research, LP plays a central role, if the amount of text devoted to the subject is a good indicator. In one of the most widely used texts, Introduction to Operations Research, 3rd. Ed., [HILL80] Hillier and Lieberman devote six of eighteen chapters to LP, and make occasional references to it in several others. If LP is so pervasive in the application and teaching of operations research, what is its particular appeal? More importantly, what is its epistemology? And, as a planning system, what precisely does it do?

Linear programming (LP) is not programming in the sense that we program computers; rather, LP selects which activities to carry out among a number of candidates, and the level of each activity. In a linear program we want to find a program (or plan) that will

$$\text{Maximize } Z = c_1x_1 + c_2x_2 + \dots + c_nx_n \qquad (2.1)$$

$$\begin{array}{ll} \text{Subject to:} & a_{11}x_1 + a_{12}x_2 + \dots + a_{1n}x_n \le b_1 \\ & a_{21}x_1 + a_{22}x_2 + \dots + a_{2n}x_n \le b_2 \\ & \quad\vdots \qquad\qquad\qquad\quad \vdots \qquad \vdots \qquad (2.2) \\ & a_{m1}x_1 + a_{m2}x_2 + \dots + a_{mn}x_n \le b_m \\ & \qquad \text{all } x_j \ge 0. \qquad (2.3) \end{array}$$

where

x_j = the amount or level of activity j;

c_j = the cost or contribution of x_j of one unit of activity j;

a_{ij} = the amount of resource i used by one unit of activity j;

b_i = the amount of resource i available;

Z = the total contribution or cost of the program of activities we select.

In the vernacular of LP, equation 2.1 is the objective, and the linear inequalities 2.2 and 2.3 are called constraints.

As with all rational models, the assertions or consequences of LP follow directly, by orderly rules of deduction, from its assumptions. In this formulation are these four, according to Hillier and Lieberman [HILL80]:

1) Proportionality: (1) the contribution to Z of x_k is $c_k x_k$, where c_k is constant, regardless of the magnitude of x_k; (2) resource usage is $a_{ik} x_k$, where the m resource usage coefficients a_{ik} do not vary with the magnitude of x_k;

2) Additivity: activities do not interact. Terms with products of variables, e.g., $3x_2x_3$, do not appear in the objectives or in the constraints;

3) Divisibility: activities can be conducted at non-integer levels;

4) Determinism: every contribution, resource level and technological coefficient is known and constant.

Based upon these postulates, widely known theorems state that optimal solutions (if any exist) occur at the convex extreme points of the solution space, that those extreme points are finite in number, and therefore, that only a finite number of steps are necessary to find an optimum. The simplex algorithm, with three decades of sophisticated refinements, is now capable of solving extremely large problems, with surprising but not clearly understood efficiency. Significant advances have also been made in the development of methods to solve planning problems when one or more of these assumptions does not hold.

To this point, we have given a very typical and very grossly understated description of the properties of linear programming. When we more closely examine the seemingly simple system defined in (2.1)-(2.3), we find that it implies another very sweeping set of assumptions:

5) Range: despite the fact that coefficients are derived from limited historical experience over the possible range of x_k, any solution is presumed to be implementable, even when one has no actual experience in that range.

6) Holism: nothing irrelevant or erroneous is included in the model, and nothing relevant is excluded.

7) Monotheism: a single objective integrates all activities into a single measure, by correctly incorporating the values of those served by the LP.

8) Causality: actions in the optimal plan cause the optimal objective value to be exactly achieved. In effect, an LP guarantees its own optimality.

We will find these assumptions far more difficult to defend when we subject LP to the criteria we gave earlier for an optimal system design.

Deciding if LP is an appropriate inquiring system for the DOD aeromedical system requires the strong systemic judgement that it is best among all alternatives. Scientific disciplines whose research create methods such as LP are silent on this issue; they disclaim that inappropriate use of a model is the fault of its developer, and not of the model itself. After all, deductive models such as LP produce knowledge rationally, through orderly rules of deduction, without any blatant contradictions, from plausible, and to be accepted, highly defensible assumptions, which are the responsibility of the model developer to select and defend. The fundamental flaw in this argument, ironically, is that there is no scientific method for selecting and defending those assumptions and hence no scientific way to prove that LP is appropriate for a given application. For that matter, since we cannot establish the appropriateness of any one scientific method, we can never settle the issue of which one is 'most appropriate'.

Correctness and comprehensiveness are two other major issues in establishing appropriateness. The sixth LP assumption above presumes that nothing irrelevant or erroneous is included, and that nothing relevant is excluded. We could add the presumption that, in addition to theory, LP computer codes operate correctly. No proof of correctness or comprehensiveness exists to assure this, although we cannot argue that they necessarily are wrong either.

That leaves the burden of proof to empirical observation. Suppose we constructed and ran an LP (and our computer software correctly solved the model), and informed one particular hospital administrator that, in the interest of the whole system, we were eliminating two stops per week at his facility that we have made for the past ten years. Verifiying the model would probably be difficult, even if it was correct, since he might feel that something is 'obviously' missing. That there may a 'perfectly rational' explanation that an LP has the peculiar property of selecting exactly as many activities as it has constraints may not convince system managers, and we would not achieve the consensus needed to verify our model. The presence of several thousand constraints and variables makes empirical verification of comprehensiveness and correctness virtually impossible, yet models of that size are routinely used in airline planning and oil refinery production applications.

There is a third issue involved in judging the appropriateness of a particular scientific method. Models must be tractable; unsolved formulations, no matter how elegant, are nothing more than restatements of the problem. LP, for example, currently does not have an adequate capability to solve large models with integer variables, yet the aeromedical system cannot move fractions of patients. The routing problems we encounter in aeromedical aircraft scheduling are said to be NP-complete, which means that beyond a certain problem size (in terms, say, of numbers of patients to be moved) we may not be able to find a computer with enough memory to solve our model, or we may not be able to wait long enough for solutions to be found.

In short, there is no theoretically defensible way to select an appropriate, or the most appropriate, inquiring system for aeromedical transportation planning. There is another problem that is at least as troublesome. No matter how appropriate (correct, comprehensive, or tractable) an inquiring system we choose, its ultimate value lies in producing knowledge that can be translated into action. In planning, we must ensure that whatever actions are planned are carried out (implemented), as planned. There are a number of dangers that this will not happen.

For the outcome planned by an LP to be realized, all system entities must behave as predicted. The planner's inability to correctly predict the behavior of system components is often attributed to faulty forecasting. What such an assessment overlooks is the failure of the LP to create the incentives for the various components to adhere to the plan. When we examine the particular kind of LP formulation that we use to model the aeromedical problem, in which objective attainment is allocated to a number of organizational subunits, we implicitly assume that the units are cooperative, or at least compelled by something not represented in the model to make contributions to the whole organization, even against their own self-interests.

We are not suggesting that we must naturally mistrust everyone, but that incentives are not easily represented.

A second implementation danger is the failure to specify in the LP the actions that should be taken in the event that changes occur that make plans infeasible. In other words, a very important part of the implementation process, is to plan for a change of plans. Through sensitivity analysis, LP models can be made adaptive, to a certain extent, to changes in problem data, but not, of course, to problems of model misspecification, at least not without the aid of something outside the model.

The third implementation danger, which stems from the whole system specification problem, is that we may be misled by LP to do, albeit very precisely, the wrong thing. Failure to specify a relevant constraint may have little or no effect on simplex calculations, but could produce completely erroneous solutions. We can also find significant ethical difficulties in implementation. As we will see, there are significant questions of tradeoffs between serving individual patients and efficient resource utilization, and of treating patients as both means and ends.

Our list of implementation dangers could go on, but one more should be mentioned. Even if someone knows an optimal solution, somehow system managers must be induced to carry out the plan. Earlier, we mentioned the incentive problem.

Another potential provblem is that managers may not have the capability to compute or recognize optimality. In the The Behavioral Theory of the Firm, Cyert and March [CYER63] argue that analytical systems such as LP miss the mark on two counts: managers do not optimize, they 'satisfice' (work toward acceptable levels of achievement), and they do so because their computational capability (their rationality) is bounded. And, Woolsey says, "Managers would rather live with a problem they cannot solve than a solution they do not understand". [WOOL80,p.5] As a result of their 'bounded rationality', managers are not necessarily motivated to find solutions unless and until they perceive that a problem exists. And, their search for solutions they can accept is often restricted to the use of relatively simple methods that consider solutions closely related to those that have worked in the past. [CYER63] In contrast to rational inquiry, the process of planning and implementation described by Cyert and March is one of small changes, which Lindbloom calls incrementalism. [LIND59]

The relationship between planning system and decision makers also portends implementation difficulties. Mason considers LP a decision-taking system. Since an LP presumes that an accounting information systems exists to specify all coefficients, that all (linear) relationships among activities are correctly specified, and that organizational values and its choice mechanism are incorporated in the objective (and implemented through the mathematical properties of constrained linear optimization models), the optimal decision an LP finds leaves the decision maker with the options of either ("rationally") taking the actions it specifies, or ("irrationally") vetoing it. With the exception of oil refinery applications, Mason could not find other examples of LP-based, broad-scale organizational planning in which normatively prescribed decisions are always followed. [MASO75] Simon [SIMO77] makes essentially the same conclusion, that normative models do not constitute a positive theory of managerial decision making. This seems to contradict the evidence that LP is widely used. Can we conclude either that organizations plan, but do not act, "rationally", since they apparently use models such as LP quite extensively, or that it is LP planning itself, epistemologically speaking, that is not rational? We will attempt to answer this important question, but only after we discuss the issue of improvement.

If selecting an appropriate inquiring system and implementing its plans are troublesome, determining the amount or degree of system improvement can be even more difficult. The essential problem is to judge if the actions taken by implementing its plans have improved the condition of those served by the system. This again is largely an ethical question, the answers to which depend upon strong systemic judgements. For example, in aeromedical transportation, we need answers to such questions as

1) Who should be served by the system?

2) What system goals are appropriate? What if the current system goals seem inappropriate?

3) To whom and to what extent should solutions be revealed? (At high wartime operating levels, the aeromedical system might be unable to move all patients, with that result that some may die. As clients, do they have the right to know this? Who decides?)

There are two paradoxes involved in the determination of improvement. The first has to do with the way we measure the amount of improvement, which in LP is through the objective function or functions. Suppose the objective of an LP must be expressed in terms of service to system clients. We can say, first of all, that there is no scientific way to establish the scope of an LP objective; mathematically, there is no way to choose between an objective that, say, minimizes the lengths of routes traveled by aeromedical aircraft, versus another that calls for maximizing the achievement of good health by the entire DOD health care system. And secondly, objectives are expressed in terms of cumulative benefit, and not in the distribution of it among clients. For the same level of objective achievement, we can provide patients widely varied quality of service, as we demonstrate in the next section.

A paradox of system objectives arises in the following way. If we can resolve the questions of the scope and distribution of achievement, then in principle we should be able to specify system objectives that express the system's intents and purposes. But precisely who should "we" be? In other words, to establish objectives (which all purposeful systems have), someone's objective must be to set objectives. How do we select such second-level objectives that resolve the first-level objective-setting problems of scope and distribution? Who, among goal-seekers, has the goal of seeking the system's goals? For the moment, let's call 'them' the decision makers. How do we measure their achievement in choosing measures of system achievement?

Any answer seems to point to a third group, whose goal is to choose the goal-setters, which requires a third set of objectives. Let us call 'them' system designers. Logically, of course, while we are at it, nothing prevents us from suggesting a fourth level, at which the principal concern is assessing the performance of those who choose the goal-setters to select the system goals. We have no convenient label for the fourth group, although we have some specific examples in mind, such as licensing bodies.

We can see, then, that the specification of objectives involves a number of complex and interrelated decisions by more than one set of decision makers. The effect of the paradox is to obscure the distinction between system roles. For example, the systems designers who choose the decision makers become decision makers themselves, and may even become beneficiaries of system performance, through the fees and salaries they receive for their performance.

The second improvement paradox also has to do with measuring system performance. Essentially, measuring performance means determining all relevant costs and benefits, which requires data to carry out a cost-benefit analysis. Typically, we use historical data on past system performance, but to what purpose? We use the data to generate the optimal plan for system activities. If the system were operating optimally, we would have one set of data reflecting this. If it were not, we would have a second set. To seek improvement implies non-optimal operation, and the existence of the second set, but not the first.

Data and plans, then, are interdependent, and hence the paradox. In the aeromedical problem, for example, we have observations of patient movement. If our assertion is correct that doctors adjust demand to system schedules, which produces non-optimal system behavior, then our demand data reflects or is biased by that behavior. That demand might differ significantly from what we would observe if the system operates according to the plans generated by our LP.

The data we utilize also depends critically on what we consider to be the 'system'. We could, for example, define it to consist of all DOD medical facilities in the US and the transportation system connecting them, and measure patient movement through the network, resource utilization, bottlenecks, etc., the stuff of typical operations research analyses. But, the movement of patients is also linked to other subsystems. More than just patient origins and destinations, hospitals are complex systems with different treatment capabilities, patient handling capacities, and staff training programs. These training programs within the hospitals affect patient movements into and out of the hospital, depending upon training case load requirements, patient treatment needs, and other factors. Patients are not homogeneous, and each entitlement program that generates patients is a candidate subsytem to be included in the system. The extent to which we include these various subsystems and the linkages between them strongly influences our choice of data to measure system performance.

So, besides the problem of utilizing data from an imperfectly functioning system to develop plans to improve that system, we have the additional problem of deciding what data is relevant, which depends upon our view of what constitutes the system. Our data collection necessarily changes as we incorporate more and more system components and linkages. And, of course, nothing systematically tells us how to do this, and we again fall back upon making strong systemic judgements to overcome these paradoxes.

Table 2.1 summarizes our discussion of LP as a suitable design methodology for the DOD domestic aeromedical trans-

TABLE 2.1

COMPARING LINEAR PROGRAMMING WITH OPTIMAL PLANNING SYSTEM DESIGN CRITERIA

Major Design Task	Critical Issues	Related Design Problems
Choice of Method	Appropriateness	Making defensible choices
	Correctness	Model verification
	Comprehensiveness	No adequate tests
	Tractibility	Modeling limitations (e.g., integrality and NP-completeness)
Implementation	Incentives	Cheating, inaction, or wrong actions
	Prediction	No satisfactory theory
	Anticipating needed changes	Forecasting Static formulation of dynamic problems
	Precision	Solving the correct problem
	Ethics	Goal appropriateness Choice of beneficiaries
	Decision support	Usurping management prerogatives
Improvement	Objectives	Scope Distribution of benefit
	Measurement	Comprehensiveness Reliability Representativeness

portation planning system. The problems associated with using LP seem to strongly argue against its use, since it seems to fail against every criterion. For that matter, it would seem that any method would likewise fail.

What we failed to do before we began to examine a specific approach was to specify what any design methodology should do. Churchman suggests that any system design should:

1) Attempt to distinguish in thought between different sets of system behavior patterns.

2) Try to estimate in thought how well each alternative set of behavior patterns will serve the system's specified goal(s).

3) Seek to communicate its thoughts to the system's decision makers in such a manner that they can convert the thoughts into corresponding actions which in fact serve the goals in the same way the design said they would.

4) Strive to avoid the necessity of repeating the thought process when faced with a similiar goal-attainment problem by delineating the steps in the process of producing a design.

5) Attempt to identify the whole relevant system and its components, so that design alternatives can be defined in terms of the design of the components and their interrelationships. [CHUR71]

The last characteristic is perhaps the most important, and in terms of designing a planning system, the first order of business. It establishes two essential tasks: identifying the system and defining design alternatives in terms of system components. The balance of this chapter will address the first, leaving the second to the next chapter.

In system identification, one of the major philosophical difficulties in describing complex phenomena, including systems, is to establish a method by which to convey the form of the phenomenon as comprehensively and as understandably as possible. In the following sections we have used Churchman's method. As Kant did in <u>Critique of</u>

Pure Reason [KANT88], Churchman sought to devise a set, or system if you will, of (non-exclusive) categories that comprehensively and collectively exhaustively convey the form of a planning system. Because the ultimate value of a social system is measured through the human needs it serves, Churchman constructed three major categories around the people who are at the center of the planner's reality: the system's clients, decision makers, and planners. The decision maker can realize changes in the system; planners conceptualize and evaluate change in terms of service to the clients, and then attempt to influence the decision makers to realize those changes which benefit the client.

The categories serve at least four purposes. First, they are intended to operate universally and a priori, and not to apply only to a specific system. Secondly, they are intended to provide the means, through a set of labels, of understanding the process of comprehending purposeful reality, of making reality intelligible. They are designed to make explicit the realistic and ethical components of design, through which a design transforms what is into what should be. And finally, they afford us the capability to communicate our description, design goals and plans in terms of the definitive categories.

From the categories, which are further subdivided into three subcategories, he devised a set of axioms on the suitable design of a planning system:

1) Every deliberately planned human action should serve a specific class of individuals called clients (and should not serve some other classes).

2) Clients are served by attempting to achieve ethically defensible goals, objectives, and ideals.

3) There should be an integrating theme (measure of performance) for client service.

4) For every deliberate action, there are decision makers who should (and others who should not) co-produce the action .

5) Decision makers co-produce actions by using appropriately resources and sets of resources called components that they can and should be allowed to use.

6) For every deliberate action, there is a co-producer of goals, objectives and ideals, which cannot and should not, be changed by decision makers, called the environment.

7) There exists a class of actions which should be planned by an appropriate group or individual.

8) Such appropriate plans ought to be implemented.

9) There exists (or ought to exist) a guarantor to prevent the disaster of erroneous plans and guarantee progress through correct plans. [CHUR79]

With this framework, we will attempt to characterize the aeromedical system through the categories, and describe how the system functions. We will conclude with our view of a major aeromedical planning problem that can be addressed with methods defined in the next chapter, and designed and tested in the remainder of the thesis.

2.2.1 **Clients**. The task of identifying the aeromedical system's clients, those whose condition is improved through the performance of the DOD aeromedical transportation system, is more complex than than it might seem. For the systems approach, it as least as important to ask who ought to be served by the system as it is to ask who is served. Determining who ought to be served involves choice, which implies criteria for client selection. Furthermore, client choice is inseparably tied to the system's purposes, which should be functions of the clients' interests.

We first attempted to construct a list of clients, the benefits they receive, and the basis for those benefits. Because the question, "Who is served?", led to the exclusion of no one, we then asked, "Who should be served?". The list remains the same, but the reasons for including each beneficiary class are more interesting.

Client	Benefit	Obligation
Patients	Treatment and the possibility of cure	DOD employment benefit
Patients' Families	Protection of family financial resources	DOD employment benefit
System Employees	Salary, intangibles such as marketable job skills	Laws governing salaries and entitlements
System Managers	Employment benefits (salary, advancement)	US Public Law
Vendors	Individual (salaries, advancement) and corporate rewards (profits, dividends)	Contracts obligating the US Treasury
Places where the system operates	Economic benefits of expenditures and tax revenues	Contracts, local and state tax laws
Nation	Increased protection through readiness	US Constitution

For active duty patients, medical care, including ancillary services such as transportation, is a benefit of employment provided by the Defense Department by law. DOD explicitly tells its employees that, since their major medical needs are provided for, there is no need for them to carry medical insurance. [DEPA82] Military employment can involve isolated or hazardous duty assignments where private care is not available, or for which the service person would have to pay inordinately high costs and insurance premiums (which usually exclude war and aviation hazards). The health benefit is not negotiable; the US soldier's right to engage in collective bargaining, job actions, and strikes is abridged under the Hatch Act. Furthermore, employee care is compulsory; e.g., certain employment standards are enforced through mandatory physical examinations.

It would seem, then, that through the conditions imposed on the employee by DOD, and its acknowledgement of its obligation to provide care, the organization legally and morally commits itself to provide medical service to those on active duty. And because free treatment after retirement is offered as a means to induce active duty members to commit themselves to military careers, retirees seem to be ethically right, as well as legally entitled, clients. Earlier we mentioned that service family members are also entitled to medical care under public law. They, too, can be subjected to some of the same circumstances in which private care is not available, which suggests that the entitlement is justified. By expanding the patient base with eligible family members, larger training programs can be supported, and since training programs achieve medical readiness, a case can be made that family members are contributors as well as beneficiaries.

Whether the system's employees, managers, and vendors should be clients is a difficult question. As long as the system requires resources, some will benefit from the economic transactions involved in obtaining and using them. It seems that an ends-means test is applicable here: economic benefits should accrue to those whose contributions (labor, goods, management) are made to achieve the system's ends, and not to exploit the system as means to their own ends. We have an example of the latter case in mind, but it first requires us to introduce another potential client group .

The people of the United States constitute a very interesting set of clients. As taxpayers, they provide the resources for defense; as citizens, they are served to the extent that military readiness achieves the goal of defending them. We are not arguing that taxpayers should necessarily be clients simply because they pay for the system. Rather, because they pay taxes to achieve a common good, they seem to be entitled to the benefits of defense, if national defense is a goal that ought to be pursued. So, the question of whether US citizens should be clients unfolds into the most critical and difficult systems question that we need to answer: Is defending its citizens an ethical national pursuit?

Economists respond by saying that, to individual citizens as consumers, defense creates a "public good", providing each citizen the same benefit, regardless of how much he contributes to it, although their microeconomic models are unable to measure the costs and values. They also argue convincingly that the macroeconomic effects of defense spending are particularly significant to all citizens, because of the magnitude of defense spending and its impact on the national economy, and because resources committed to defense cannot be used for other societal purposes. Thucydides raised this issue 2400 years ago:

> What good is building a magnificent machine of destruction to protect one's valued way of life if the process is so costly that it wastes that way of life as surely as if one's enemy had triumphed on the battlefield?

But, beyond establishing how their economic interests are effected, economic theory does not address the ethical question of whether all US citizens should be clients.

From a systems standpoint we would argue that we cannot answer the question of whether any individual or group should be clients without considering the interests of all other potential clients. As we will demonstrate in our discussion of measures of performance, service to each individual patient is inextricably tied to serving all patients. In turn, because patient service uses public resources, the interests of the public are also involved, which suggests that they must be considered as clients. But the unfolding of the client list does not stop with the people of the US, since US defense pursuits have significant effects on other nations. Our list should also have included other nations who benefit through mutual defense treaties, and perhaps even all nations, to the extent that major conflicts are deterred. (The negative effects of system failure could also be used to argue the same point).

The client question, then, continues to unfold until utimately we must consider whether the interests of all human beings are effected by a US defense subsystem, thereby making them clients. In reality, we would argue that the planner's main client is always humanity, because linkages between all client groups always exist and cannot be ignored. These linkages are particularly important to our system description, because they characterize the inherently conflicting interests of the various client groups. And, as we will show in the next section on system purposes, some conflicts among clients are unresolvable.

Ethically, whether humanity should be the main client is not as apparent, because it involves the question of the system's purpose, which we have not yet discussed. But this is ultimately a practical issue: How large a client group can we adequately represent in our model? In our judgement, we can to some extent represent patients, system members, and US citizens, but we have to exclude all others, realizing fully the sacrifice in comprehensiveness.

With this in mind, we can return to a group we deferred judgement on earlier, the system's employees, managers, and suppliers. If we must have an aeromedical system, we should compensate those who provide its resources and management. The system should not serve those who perpetuate the system for their own advantage. If this and other defense system components were perfectly designed, they would not exist at all, which would constitute the optimum level of client service. This argues strongly against such practices as maintaining constant budget levels, which seem directed more toward making the management task easier than achieving an end that ought to be pursued.

Before we begin our discussion of purposes, we should mention a recurring problem in relating client benefit to system performance. We have identified and chosen clients as members of large classes, and it is easiest to rate aggregate service to all class members. Even if the aggregate benefits are large, the distribution of them to each individual client is also important. This seems particularly true in the case of medical care, where attention to individual needs is extremely important. We will illustrate this in our discussion of measures of performance.

2.2.2 **Purposes**. In our discussion of the clients of the aeromedical transportation system, client choice was based on obligations to serving them, a narrow and more passive view of why they ought to be served. As planners and decision makers, we actively intend for future states of the world to occur that will benefit clients (just as we do not want other non-beneficial or harmful states to occur). Pragmatically, these intentions must be achievable, through appropriate decisions and actions, and within the limited system resources we have available. These realizable intentions are what we call system purposes.

This section examines the purposes of the aeromedical transportation system. Purposes have both realistic and ethical components, which need to be identified. The category of purpose unfolds further into subcategories that provide a convenient classification scheme, and more importantly, a means of explicitly linking our client service intentions with measurements of how well we achieve those intentions. Finally, we will look at prospects for changing the way in which purposes are pursued, since the essential concern of the systems approach is ultimately practical: designing a planning system we can reasonably expect to implement.

In most studies, this section would be labled "objectives". Every teleological (purposeful) social system has them, although we often find them pursuing multiple, conflicting, often vaguely stated, or even misleading or unstated, purposes. Even when carefully and officially articulated, they are not necessarily pursued in an equally coherent fashion. Recalling our discussion of the DOD readiness and benefit objectives, the Defense Resource Management Study concluded that

> DOD responsibility for the benefit mission is assigned by Title 10, United States Code, Chapter 55. DOD Directive 5136.1 delegates to the Assistant Secretary of Defense (Health Affairs) the authority to " ... issue ... regulations .. to fulfill the Secretary of Defense responsibility to administer ..." the benefit mission.
>
> Army, Navy and Air Force regulations that assign missions and functions to the respective Surgeons General and to commanders of service medical commands and hospitals fail to assign explicitly the benefit mission. In certain instances the regulations, by inference, implicitly assign the benefit mission. But nowhere in service regulations was the DRMS able to find explicit assignment of the benefit mission to health care managers.
>
> The absence of an explicit health benefit mission leads to an unnecessarily convoluted train of logic in the justification of resources needed to accomplish the benefit mission. [DRMS79,p.82]

This illustrates one important distinction among system purposes: what is stated versus what is real (and often hidden or unstated). A major problem or fallacy associated with identifying purposes is to overemphasize obvious purposes. One common statement of purpose for the aeromedical transportation system is to move patients as rapidly as possible. But what the system seems to be really trying to do is to move patients so that

1) Patients recover more quickly (for their benefit);
2) Patient comfort is increased (for their benefit);
3) They return to duty sooner (increasing readiness);
4) The planning system is exercised (maintaining readiness).

Other purposes are expressly stated, such as maintaining a level of system operation sufficient to keep all system employees and decision makers trained in their respective roles, when other unstated or hidden purposes may be involved. The conflict is between what the system is really trying to accomplish and what it is willing to reveal about those aims. The organization sets and meets training goals to achieve the readiness objective, and annually requests resources resources for that purpose. Alternatively, we could assert that formulating the same training goals every year results in constant annual budget proposals, which in turn avoids all the additional work and possible scrutiny involved in changing them. [WILD75] Unless the request made ten years ago correctly related training needs to resource requirements, which suggests either that system managers possessed an exact causal model of the training process or made particularly fortuitous guesses, the circumstantial evidence is that minimizing effort and controversy is one underlying purpose served.

Two further distinctions can be made regarding system purposes. First, among multiple goals, an organization may knowingly sacrifice or ignore some goals in order to pursue others, and its managers may or may not realize the opportunity costs involved. Somehow, organizations (overtly or not) order, rank, classify, or sequentially meet purposes in order of importance.

Secondly, in our statement of the category of purpose, we distinguished between what is and what ought to be pursued, what we might label the differences between the real and legitimate purposes of the system. For example, system managers must frequently explain past planning decisions that do not seem to achieve an acceptable level of productivity. When the Government Accounting Office and the Air Force Audit Agency both held that some flights were dispatched at considerable expense to serve an unacceptably small number of patients [DEPA78b], system managers responded by imposing restrictions (such as requiring that a minimum number of patients be served in order to launch an air-craft) that had the apparent result of increased effic-iency, even though they realized that service would be sacrificed. The externally imposed efficiency target was achieved, but at the expense of more legitimate goals.

It is very useful to discuss organizational purposes by unfolding them into a series of subcategories which we earlier called goals, objectives, and ideals. In their usual definitions, they differ by time horizon, degrees of attainment or both. We define them somewhat differently.

We define goals as perceptions of a planner who takes most aspects of a plan to be given. "Targets" might be a preferable synonym. Goals make the relationships among clients, purposes and performance measures readily apparent. Apparent also are data requirements (compiled from records of past system behavior and careful recording of current events), and the decision outcomes that will increase or decrease attainment. In goal planning, the planner views virtually everything as unchangeable.

In objective planning, which we illustrated in our LP discussion, the "givens" are embedded in a larger framework of constraints. LP emphasizes bounding a planning problem to determine feasible and realistic alternatives. Unified measures of relevant costs and benefits are used to find the "best" point within a feasible range of choice. One currently emphasized aeromedical transportation objective is to reduce patient overnight enroute stops, given all resource limitations, patient handling rules, and other factors. The desired outcome is not to reduce overnight stops to some acceptable level, but to achieve the lowest level possible, given all constraints.

Ideals are purposes that could hold if feasibility and reality constraints were removed or ignored. [CHUR79] The principal concern of ideal planning is whole system improvement, which necessarily must be measured in terms of system ideals. One ideal might be a perfectly designed aeromedical system, but that would depend on designing all defense systems perfectly. The only reality that would permit such a perfect design is one in which there is no war, assuming a perfectly functioning defense produced that end. Since most of us seem inclined to accept war as an inevitable reality, there is little likelihood of carrying out ideal defense planning if we can't reject war as a given. But beyond our own cognitive limitations, there may be a more fundamental problem: valid ideals may not exist. Churchman argues out that no completely defensible ideals have ever been shown. Even if we could specify one, we can inevitably find another that conflicts with it; ideals serving all of humanity necessarily conflict with those directed toward each individual. [CHUR79]

Beyond talking about purposes, we needed to identify some representative ones to incorporate in our system model. But as Karl Borch astutely observed, "In real life, it may well be more difficult to decide what one should maximize, than to carry out the actual computations in the maximizing process." [BORC70,p.5] Our bias was to focus on real purposes the organization seemed intent on pursuing that, in our judgement, they ought to pursue.

In March, 1983, the Military Airlift Command Surgeon General's staff and the parent organization, the 375th Aeromedical Airlift Wing, jointly studied a number of current planning problems. [DEPA83] From their final study report, we have derived the following implied or explicitly stated goals and objectives:

Concern	Goal	Objective
36 per cent of all patients remained overnight at least once in CY82.	None stated; "0 per cent is unresonable; 36 per cent is too high."	Minimize the number of patients required to remain overnight (RON).
Demand for patient tansfers increased approximately 50 % from calendar year 1978 to 1982.	No increase in the number of patients required to RON.	Accomodate all new movement requirements without more employees or aircraft.
40 CONUS medical treatment facilities (MTFs) without airfields must routinely drive patients to meet flights.	Add 21 stops to those routinely included in the daily schedule.	Eliminate trips by ground transportation of more than one hour.
Provide initial pilot checkouts and advanced training.	Fly 120 hours on local flights at the central base.	Maintain a maximum level of qualified flight crews.

Their concern for patient welfare is apparent. In addition to these purposes, the study also indicated that, despite significant increases in the number of patients transferred, system managers wanted to continue to provide direct point-to-point service to patients in urgent and priority categories or who have special needs requiring such service. Recognizing that additional daily flights would not be feasible without more crews, which Congress must approve and fund, they wanted to achieve these ends without changing the fixed, six-mission-per-day schedule.

We were particularly interested in the lack of any economic goals or objectives, such as minimizing fuel use. Fuel economy is an area of concern to DOD, since defense is the dominant federal sector fuel consumer (97 per cent in FY80), and aircraft operations used over 65 per cent of all DOD petroleum in FY80. [DEPA81] As we note later in discussing system resources, constant fuel prices are partly responsible. Another explanation is that maximum, rather than optimum, fuel loads are carried, in order to have maximum flexibility to respond to changing demands, particularly cases requiring immediate pickup and delivery.

We also noted that the annual budget is just about exactly exhausted each year. The reason is that the standard practice dictated by higher authorities is to treat the budget as an amount not to be exceeded, and not to be underused by more than two per cent. Not only that, while manual flight scheduling methods might attempt to minimize the length of some flights for patient comfort and other reasons, schedulers lacked any training in or automated support for systematic route optimization. It was apparent that purposes benefiting the taxpayer-client (who provide the annual resource budget) were not actively pursued, even though we argued in the last section that they ought to be. Changing long-standing institutional practices such as pricing fuel uniformly and exhausting budgets completely, which are imposed by higher authorities outside the organization, would be difficult, particularly since a concensus among all working-level organizations (for many of whom preserving these practices would be to their advantage) would probably be required.

While we do not propose any aeromedical transportation system ideals, we can at least adopt the ideal planning practice of looking beyond whatever "givens" we can, attempting to determine what level of service the system could achieve if some "realities" were different. Aeromedical planners, for example, assume that facilities where patients can be accomodated overnight are fixed both in numbers and locations. But in assuming such realities, they may overlook other possibilities. Instead of questioning facility quantity or mobility, it might be useful to determine if they should be used at all. Their existence may influence planners to think that they are absolutely necessary, when more aircraft might completely eliminate the need for them, and lead to significant whole system improvement. As another example, self-imposed scheduling restrictions, based on such perceived realities as maintenance contracts and fixed fleet size, dictate the same number of missions each day, even though patient demand varies widely over different days of the week. Varying the number of daily missions may lead to more responsive service.

Before we conclude our discussion of system purposes, we need to examine the implementation potential of objective planning. Objective planning approaches to public sector problems involving optimization have not enjoyed much success, because as Lowe and Moryadas observe,

> In affluent societies and in normal times, the price of nonoptimal behavior may be tolerable. However, when a ... national emergency exists, efforts are launched with a view to promoting efficiency and curtailing needless expenditures.[LOWE75,p.283]

We will have more to say in our discussion of the system's environment about the lack of incentives to pursue objectives optimally. Our concern here is pragmatic: system decision makers have not developed optimization practices or habits, as evidenced by the current lack of capability to search for better or the best solutions, and, more significantly, they apparently lack interest in finding one. In this study, we found that the underlying subsystem problems are so inherently difficult, that unless the organization develops and uses (or at least tests) a methodology able to find more efficient allocations of its resources, there is little likelihood that it would be able to after a crisis begins. Improving the planning system seems to us to be a purpose system managers legitimately ought to pursue. (And, we recognize immediately the paradox of ourselves as planners making decisions that prescribe motives for the system).

2.2.3 Measures of Performance. On the surface, the DOD medical transportation problem seems very simple. Whenever a patient and the medical service he or she needs are physically separated in space, the service has no value until the two are brought together, and the service attains its greatest value when it is administered when it is needed. Trans-portation increases both the time and place utility of med-ical care by moving patients to treatment, or vice versa, with minimum delay.

Medical transportation planners use information on where patients need to go and when they need to be there to devise aircraft schedules that will deliver patients to their destinations as quickly as possible. Determining how well the planning function performs requires us to specify an adequate measure of how well the system meets the needs of its clients. This turns out to be quite difficult because of inherent conflicts that exist in most, if not all, mass transportation systems. The divergent interests of various client groups create conflicts. Minimizing aeromedical transportation costs would best serve the interests of taxpayer-citizen clients, while minimizing delays and discomfort is more important to patients. To choose one measure (or even several) is to decide how to allocate system costs and benefits among groups, and among individual clients. In the presence of multiple and conflicting goals and objectives, selecting one 'best' measure(s) is virtually impossible. In this section we illustrate a number of inherent conflicts. We conclude that two standards ought to be considered, minimizing patient transfer delays and total transportation time.

The central concern of economics is allocating scarce resources to satisfy human wants. However, economics has had remarkably little to say about transportation. Usually, economics separates the production and consumption of goods and services. As a productive process, transportation services transport, rather than transform, commodities or people. Unusually, (though not uniquely), transportation produces a service that requires the consumer to supply one of the input factors (his time) (for which there is an opportunity cost that may be useful in assessing the value of the service) which greatly complicates assessing costs and benefits to consumer and producer. The DOD case is even more difficult, because the producer, not the consumer, pays for the service!

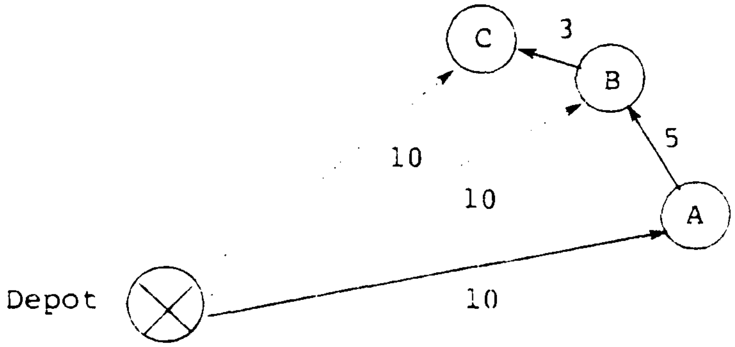


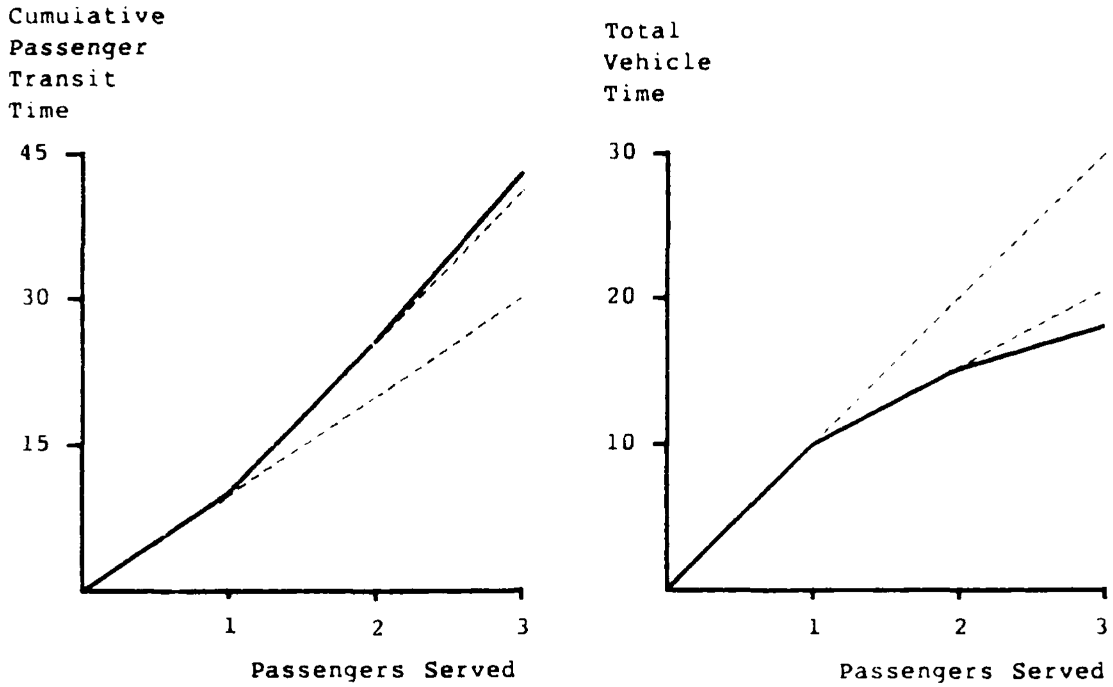
Figure 2.2. Multi-user transportation service example.

To illustrate the performance measurement problem, consider the following multiple user transportation service example. Suppose we have three passengers located at the depot, one vehicle, and the route with travel times annotated in Figure 2.2. Points A, B, and C represent the respective stops of the passengers. If the vehicle's route is from the depot to the points in the order suggested by the arrows, or arcs, then three trips are 'produced':

TRIPS	TRIP DURATION (Minutes)	MARGINAL COST (Vehicle Time, in Minutes)
1 (Depot-->A)	10	10
2 (Depot-->A-->B)	15	5
3 (Depot-->A-->B-->C)	18	3

In terms of vehicle time and passenger riding time, we observe two differing measures of the productivity of the route. Marginal vehicle time decreases with each trip, while the marginal cumulative passenger transit time increases; the amounts depend on the particular sequence.

If, in this example, the direct routes from the depot to points B and C required ten minutes of vehicle travel time each, then for the route above, the first passenger spends no time in excess of the time required for a direct trip. But, the second and third passengers endure an additional five and eight minutes respectively of excess riding time over direct deliveries. This illustrates a dilemma common to mass transportation systems known as



Fiȝure 2.3. Vehicle routing productivity.

congestion, in which, under certain conditions, one or more passengers experience excess riding time as more passengers are served. There are two major problems created by the congestion phenomenon: determining the cost of excess riding time congestion causes, and finding an appropriate measure that will choose, from a set of prospective routes, one that provides the best service to all passengers.

Suppose in our three-passenger problem, the depot and destinations are arrayed in the Euclidian plane of unit squares in Figure 2.4. The point-to-point distances in Table 2.2 result. We can define two types of routes, open and closed tours, as those that depart the depot, visit each destination once (and only once) and terminate at the last destination or at the depot respectively. The tree in Figure 2.5 enumerates all possible open tours; closed tours have an additional segment from the last stop to the depot.

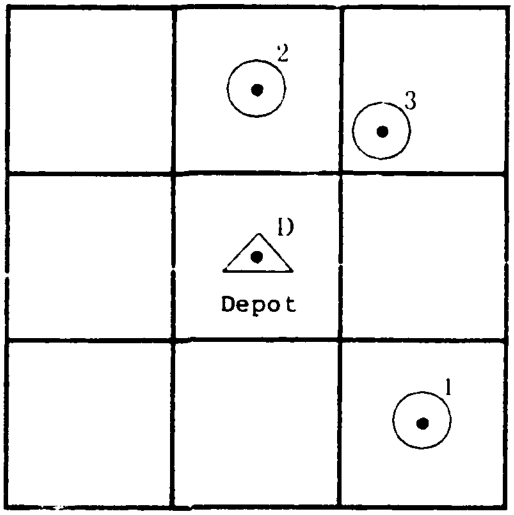


Figure 2.4. Three-passenger transportation example.

TABLE 2.2

POINT-TO-POINT DISTANCES

	TO	D	1	2	3
FROM	D	-	1.4	1.1	1.0
	1	1.4	-	2.3	1.8
	2	1.1	2.3	-	.8
	3	1.0	1.8	0.8	-

As we can see in Table 2.3, route IV is the minimal vehicle time open tour, one of two minimal time closed

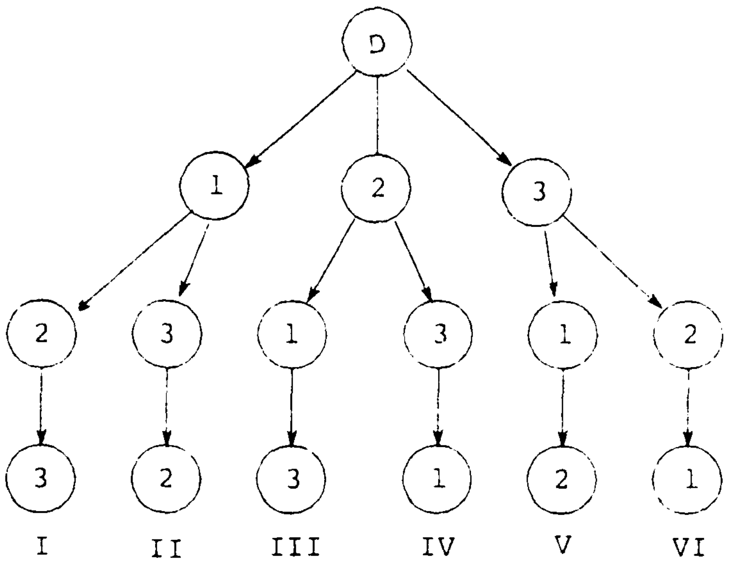


Figure 2.5. All three-passenger delivery routes.

tours (that are equivalent in length, and exactly opposite in order), and the minimum total passenger service time tour. But, note that route III, the longest open tour, one of two maximum length closed tours, and the maximum total service time tour, affords service at least as direct as route IV to passenger 2, and more direct service to passenger 1. For comparison, Table 2.3 gives the characteristics of three separate round trips as route VII.

Besides separate round trips and tours, a third possibility exists that combines these two types: multiple trips with each serving one or more passengers. Two trips, Depot-->3-->2-->Depot, and Depot-->1-->Depot, require 5.7 units of total vehicle time, .6 more than route IV. For that 11.76 per cent increase, cumulative on-board time is

TABLE 2.3

ROUTING MEASURES OF PERFORMANCE

	Route						
Measures	I	II	III	IV	V	VI	VII
Closed Tour Length	5.1	5.5	6.2	5.1	6.2	5.5	7.0
Open Tour Length	4.0	4.5	5.2	3.7*	5.1	4.1	N/A
Delivery Times for:							
Passenger 1	1.4*	1.4*	3.4	3.7	2.8	4.1	1.4*
Passenger 2	4.0	3.7	1.1*	1.1*	5.1	1.8	1.1*
Passenger 3	3.2	4.5	5.2	1.9	1.0*	1.0*	1.0*
Cumulative Time On-board	8.6	9.6	9.7	6.7*	8.9	6.9	3.5*
Largest Direct Delivery Deviation	2.9	3.5	4.2	2.3*	3.3	2.7	0.0*

reduced to 4.2 units, only .7 above direct delivery, and 2.5 units (69.56 per cent) less than on route IV.

This simple example illustrates two important points. First, no measure can be the best for both vehicle utilization and individual passenger service. In some instances, the longest route provides equal or better service to all passengers than the shortest. Secondly, a strategy that searches for only one route type (tours, for instance) may overlook routing solutions that are only moderately more expensive in terms of vehicle operation, but provide much better service to individual passengers. In all instances except direct delivery, some passengers necessarily endure excess riding time.

Another example more closely related to patient transportation reinforces these conclusions. In Figure 2.6 we depict what we define in Chapter 4 as the mixed service problem, which requires stops to both pick up and deliver each passenger. A node labeled +n designates the origin of passenger n, and -n his destination. Observing the logical restriction that we must visit a passenger's origin before his destination, 90 feasible closed tours can be constructed, versus 720 (6!) without the order restriction.

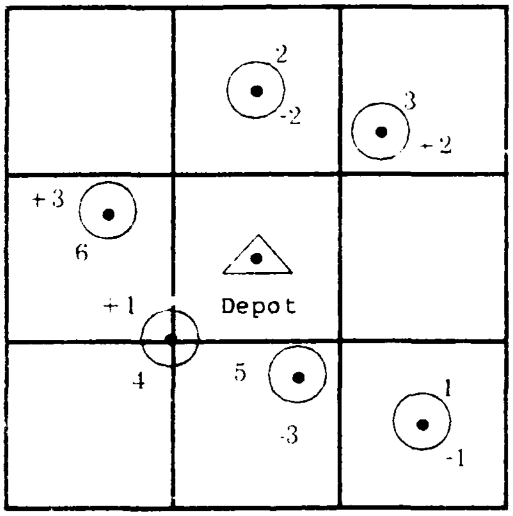


Figure 2.6. Three-patient mixed service example.

Six routes minimize total passenger riding time, while two minimize total vehicle time. One of those two, route VII, comes extremely close to minimizing all criteria.

Route	Sequence
I	D-->+1-->-1-->+2-->-2-->+3-->-3-->D
II	D-->+1-->-1-->+3-->-3-->+2-->-2-->D
III	D-->+2-->-2-->+1-->-1-->+3-->-3-->D
IV	D-->+2-->-2-->+3-->-3-->+1-->-1-->D
V	D-->+3-->-3-->+1-->-1-->+2-->-2-->D
VI	D-->+3-->-3-->+2-->-2-->+1-->-1-->D
VII	D-->+2-->-2-->+3-->+1-->-3-->-1-->D

TABLE 2.4

MIXED-SERVICE EXAMPLE DISTANCES

	TO	D	1	2	3	4	5	6
FROM	D	-	1.4	1.1	1.0	.7	.7	.9
(-1)	1	1.4	-	2.3	1.8	1.5	.8	2.2
(-2)	2	1.1	2.3	-	.8	1.6	1.8	1.2
(+2)	3	1.0	1.8	0.8	-	1.7	1.5	1.6
(+1)	4	.7	1.5	1.6	1.7	-	.8	.9
(-3)	5	.7	.8	1.8	1.5	.8	-	1.6
(+3)	6	.9	2.2	1.2	1.6	.9	1.6	-

TABLE 2.5

MIXED SERVICE ROUTING MEASURES OF PERFORMANCE

	Route						
Measures	I	II	III	IV	V	VI	VII
Closed Tour Length	8.3	8.0	9.4	8.3	8.5	9.3	6.9*
Open Tour Length	7.6	6.9	8.7	6.9	7.4	7.9	5.5*
Delivery Times for:							
Passenger 1	1.5*	1.5*	1.5*	1.5*	1.5*	1.5*	1.6
Passenger 2	.8*	.8*	.8*	.8*	.8*	.8*	.8*
Passenger 3	1.6*	1.6*	1.6*	1.6*	1.6*	1.6*	1.7
Cumulative Time On-board	3.9*	3.9*	3.9*	3.9*	3.9*	3.9*	4.1
Largest Direct Delivery Deviation	0.0*	0.0*	0.0*	0.0*	0.0*	0.0*	0.1

Heuristics that do not maximize or minimize a measure of performance can select particularly unsuitable routings. A simple rule-of-thumb, making all pickups first, and then all deliveries, finds the worst solution using total passenger time as the criterion. This suggests two attractive properties of optimization approaches. First, regardless of which criterion is chosen, there would be no doubt as to how 'far' from optimality a heuristically-derived solution might be. And, if the best solutions can be found with respect to several criteria, techniques exist that can aid in the search for compromise solutions, with the direct involvement of the decision maker.

The problem, even with optimization techniques, as these simple examples show, is that, depending upon the criteria chosen, widely different solutions may optimize those criteria. It is precisely the choice of the criterion and not whether we can optimize it that causes the most difficulty. The critical difference in service time criteria, for example, depends upon whether we include vehicle, total or individual passenger service time, or a combination, in the criteria, and what other stipulations we impose (e.g., solutions must be tours). If we choose vehicle time exclusively, in the transportation service provider's interest, routes should minimize vehicle time, if, as is often the case, that is the major determinant of factor costs. Since factor costs on the provider's side should be calculable using appropriate resource accounting data, this would not seem a difficult measure to implement.

Criteria based upon the interests of the service user are more difficult to derive. Specifically, we need to be able to determine the factor cost of the patient's time and a method to handle tradeoffs among users. Studies have inconclusively attempted to impute the value of travel time, because of fundamentally difficult problems. [LOWE75]

First, there is no clear way to associate labor cost and travel time. Military patients are salaried (not self-employed, nor employed by another organization) and therefore are traveling on their employer's time, not their own. There is no opportunity cost to them, no liesure versus labor choice. From their employer's viewpoint, there is a cost associated with the loss of productivity during treatment, including the time they spend in-transit. However, there is considerable reluctance to make service a function of relative importance -- generals before privates, if you will -- a practice that would create major ethical problems in equitably distributing benefits.

There are related problems with a salary as factor cost approach. Individual value to the organization is difficult to assess, both in general, and specifically, due to the practice of paying DOD employees within very broad categories (primarily by military rank and longevity).

Other client classes (retirees and dependents) cannot be measured by employee factor cost means, since their travel time has no meaning in terms of lost DOD productivity.

Market choice methods do not seem applicable. Patients do not make the kinds of assesments that, say, commuters do when they can evaluate the speed, convenience, comfort and costs of alternative travel modes, factors which have permitted valuing transit time in other planning contexts. Although there may be instances where travel represents a particularly valuable alternative, as it does to retired dependents who might not receive any free care unless they travel, we could not visualize a workable market-choice approach to factor costing.

Another possibility is to ignore the question of cost, and examine just the sequencing aspect of patient movement. That is, we could view an aeromedical aircraft as a job-shop processor, and patient movements as a series of jobs to be processed. A flight is, after all, a sequence of processing steps, each with a finite length, starting point, and ending point. The fact is that scheduling theory does not value the completion of individual jobs, and instead measures the value of any sequence in aggregate terms. In schedule-theoretic measures, a sequence would be evaluated to determine how many patients arrived within the prescribed criterion for delivery after being picked up, or the number of days the latest patient arrived at his destination, but not for the cost or impact on each individual patient.

For the system purposes we proposed earlier to represent in our system model, we propose two simple measures. The primary objective should be to minimize patient delays enroute, which we can measure by summing the number of overnight delays patients experience after being picked up. The principal beneficiary is the patient, although this would also reduce overnight facility operating cost. And, this measure favors direct delivery, which is particularly important in wartime operation.

Whenever possible, and without increasing the number of overnight stays, planners should attempt to find minimum cost routes, that lower aircraft operating costs and move patients, at least in aggregate terms, to their destinations more quickly. While this seems primarily oriented toward taxpayer-client benefit, the wartime need for maximum utilization of aircraft would be served by having planners systematically finding shortest routes.

We propose that the concepts of multiple objective (MOLP) and goal linear programming (GP) be merged; that the original multiple goals, and not derivative goal targets, be optimized, as in MOLP, but where the primary objective of minimizing overnight stays preempts the cost minimization.

2.2.4 Environment. In the systems literature, identifying a system's environment critically depends on determining whether the system is open or closed. In open systems, the lack of a definable boundary implies that no environment exists, while the environment of a closed system consists of everything that lies outside boundaries enclosing the system. Beyond those boundaries, beyond the control of the system's decision makers, there are "things" that determine or influence system performance, and hence, the effectiveness of the system relative to the client.

In operations research problems, environmental elements are often conveniently labeled "fixed" or "given". They imply the existence of other entities within the boundaries that can be controlled, which we have defined as "resources". There seems to be little disagreement that social systems have both environments and resources. Since these elements bound the actions that decision makers may take, the treatment of the environment is typically a search for the the "correct" boundaries (geographic, organizational, budgetary, and technological, to name a few) that constrain action.

It would seem a fairly simple matter to decide if anything is either a resource or an element in the environment of a system. By asking the question, "Does *it* matter to clients?", we could create two exclusive classes. By then partitioning the one containing resources and environmental elements by asking the question, "Can we control *it*?", the identification problem would be solved.

In the aeromedical transportation problem, aircraft are resources by this definition, since they obviously matter to and are controlled by the organization. To obtain more, which the organization currently wants to do, requires Congressional budgetary authorization and appropriation, which lie outside the aeromedical organizational boundary. Since the number of aircraft (fleet size) matters, but cannot be controlled, fleet size is in the environment.

This classification scheme fits well with objective-planning methods such as LP. "Given" the maximum fleet size of eleven aircraft in the domestic system, we could combine these resource levels, and coefficients specifying environmental limits to our using them, into constraint expressions that form the mathematical boundaries of our linear programming model. Planners could then directly determine the feasibility of any action by testing whether or not it lies within these feasible boundaries.

In effect, then, these descriptors of environmental entities that bound action create a one-to-one correspondence between restrictions or limits we observe or experience in the real world and constraints in the LP. There are many entities over which we have little or no control, but which matter, such as maximum aircraft speed, patient movement demand, and aircraft patient-carrying capacity. Once we have specified all important correspondences, our LP would provide the analytical means to control the use of our resources, within the amounts we have available and the limits of the technology governing their use.

But is the process as simple as it sounds? The beginning of the chapter described the DOD aeromedical system as part of a series of successively larger and more encompassing organizational systems. Precisely which one do we choose to study? Should organizational entities be considered system entities? And perhaps most importantly, why should we be concerned about the choice?

Unfortunately, there are no completely defensible answers to these questions. Any answers would obviously involve strong systemic judgements. We can say, however, that there are a number of dangers in ignoring them:

1) Assuming that what we have observed is all that is relevant may lead to serious planning miscalculations.

2) Failing to account for the effect of linkages between system components and the environment may cause planners to make decisions that seriously reduce service to clients, or waste resources unnecessarily.

3) Taking something as "fixed", unchangeable, or both, will not allow us to move beyond objective to ideal planning.

In the case of the first danger, something potentially useful may be absent from the system's environment that may be highly useful. In the aeromedical system, for example, no pricing mechanism forces decision makers or service users to choose among alternatives on the basis of cost to prevent uneconomical operations. There is an even more fundamental reason for considering this problem of relevance and comprehensiveness. As we said earlier, LP models contain very strong assumptions about the nature of reality. One of them is that an LP model contains everything relevant to a problem, and nothing irrelevant or erroneous. We said that plans generated by an LP should specify where every entity in a system should be and when. Because our ability to include all entities and their alternate paths through time and space is limited, we conceded that somehow we have to limit the scope of the planning problem. The major planning question, then, is ultimately a highly practical one: "How large a problem are we capable of modeling?" In attempting to capture the most salient features of the aeromedical system's environment, that question becomes: "How much of reality can we, and should we, incorporate in our model?"

In the second case, when we attempt to solve a particular system problem, we have to be aware that it can't be solved on its own basis without possibly creating other problems, because of linkages that exist among system entities. For example, suppose we presumed that one major problem was a lack of aircraft. The intuitively appealing (and typically adopted) answer would be to buy more. But as Graham shows in a series of very clever examples [GRAH79], under certain conditions, more or faster vehicles can actually degrade system performance because of precedence relationships that exist between system entities. Churchman calls this the environmental fallacy: If "x" is "too low" or "too high", we have to resist the impulse to change "x" without adequately taking into account unintended and potentially serious side effects.

The third danger is perhaps the most serious. Even when attempting to make the case that more aircraft are needed, we noticed that the organization assumed that traditional scheduling practices must continue, which may be a more significant source of their service problems than fleet size. They never questioned such givens as that the system ought to continue to exist, that the right clients were being served, with the right resources, and so forth. By failing to identify all relevant constraints, and then look beyond them for solutions that are, by current standards, infeasible, there is little hope for progress.

Selecting those things which constitute the environment is particularly important in defining the data requirements of our study. The word data itself is derived from the Latin for "that which is given". As we have already noted in our discussion of the paradox of system data, the combination of our actions and environmental factors over which we have no control will jointly determine outcomes that will generate data different from the case where we take a different (or no) action. This suggests that we must temper our reliance on historical data with a judgement as to how well it can forecast the performance of the system we are very likely going to change.

The environment we select is interreálted with our choice of the system's decision makers. Decisions are taken for the purpose of specifying action, but actions are restricted to those things over which the decision maker has control. In ideal planning, the principal concern is the decision maker's *Weltenshaaung*, or world view, and what it contains in terms of perceived environmental realities, and the values he places on outcomes that are co-produced by the environment, his decisions, and system actions.

In the remainder of this section, we will discuss what we have observed to be particularly important elements of the aeromedical system environment. In particular, we will indicate the extent to which they might be changed and the dangers attendant in doing so or failing to do so.

One aid to identifying a system's environment is to determine what is <u>not</u> "out there". The absence in the environment of the DOD domestic aeromedical system of a number of influences is significant. The DOD system is unique; there are no other organizations like it. That means that there are no successful role models to emulate, nor unsuccessful examples to learn from either. As one practical consequence, uniqueness means that system vendors must resort to expensive, custom fabrication to provide operating equipment and supplies.

In this monopoly-monopsony relationship between a single server and a single collective customer, there is no competition on either the supply or demand side. The system does not compete with private-sector, market-oriented organizations. And, funding for DOD aeromedical service is supplied by the provider, not the user. (Only in rare circumstances are the patients or their sponsoring organizations charged for their transportation). There aren't any pricing cues from either market or internal pricing systems.[16] Military hospitals do have budgets that can be used to contract for services at nearby civilian hospitals for certain patients, as an alternative to moving them to another military facility. However, aeomedical service is a cost-free alternative, particularly when contractual services budgets are limited or exhausted.

Furthermore, the organization is compelled to use all of its annual obligational authority (expressed below in flying hour units). Despite the fact that DOD operations, under the auspices of the Program-Planning-Budgeting System

(PPBS), are supposed to be ends or missions oriented, the annual program is both justified and managed in terms of resources. Where the program is supposedly driven by the combination of wartime training and patient service needs, it is guaged by the means-oriented measure of flying hours used. Assuming that, since all hours were used, all training and patient service tasks were met (which other internal reporting systems confirm), we can probably conclude that the mission was accomplished each year. What is lacking, however, is any idea of how much less, in resources consumed, might have been used to reach the same end. That is not to condemn the organization, since it must use the PPBS process, which Wildavsky asserts never has and probably never will be able to produce ends-oriented measures of performance. [WILD75]

Annual CONUS Aeromedical Flying Hour Programs

Fiscal Year	Authorized	Used	Percentage
1977	18216	17900	98.1
1978	18216	18250	100.2
1979	18216	17983	98.7
1980	18216	18093	99.3
1981	18216	17513	96.1

Source: [JONE82]

There is no compelling reason to economize. No board of directors or stockholders demand sound performance measurable in terms of profits or return on investment. There are no financial rewards to managerial initiatives to reduce costs or budgets. In fact, there are strong disincentives. Slack capacity provides the means to respond to unforseen contingencies (and makes planning easier); maintaining slack becomes an important goal. Annual budget level and organizational size are closely correlated; cost reductions or savings can lead to loss of corporate size and power. Continued Congressional funding means continued existence; funding reductions, then, are the organization's most serious environmental threat.

So we find a curious twist to the environmental identification problem. It is not enough that we simply ensure that we not overlook entities which do exist; we need to identify important environmental forces that are absent. Not finding mechanisms that force the aeromedical organization to operate effectively does not mean they should not be compelled to operate effectively and efficiently. Rather, it means that creating the required incentive-producing mechanisms is part of the design problem, which, in effect says that "designing" the enviroment is a necessary planning problem.

A number of exogenous, or at best only partially controllable, factors constrain the operation of the system. Among the more important ones are patient demand, fleet size, rules and regulations, and aircraft operating limitations. As we discuss each of these we will emphasize the way in which they affect system performance, their impact on the planning problem, and past management actions and future prospects to mitigate them.

Patient demand is determined by a number of factors.

1) At each facility, budget levels and use, new construction and closures, and the creation and elimination of treatment programs all significantly affect the need to send patients to other hospitals for care.

2) Changes in beneficiary population size and location, entitlements, and program policies also change demand patterns. For example, as dependents and retirees have been required to pay a larger share of CHAMPUS costs, more have elected to transfer rather than pay.

3) Demand growth has steadily increased 4--5 per cent for the past six years. [JONE82]

4) Emergency movements increased significantly during one five-year study period, particularly in cases with neonatal complications, burns and neurological problems; that they had even survived to be moved was due largely to significant advances in treatment methods. [JOHN76]

Medical needs are inherently stochastic, and despite some control measures, such as ASMRO regulation and some treatment advance scheduling, it cannot be completely controlled or even precisely estimated. Patient transportation demands are therefore handled on a demand-responsive, rather than a fixed-schedule, basis. Where some control is possible, despite evidence of the adverse effects of service parochialism and the lack of strict criteria to minimize interregional transfers, no measures have been taken to eliminate these practices during the time we have studied the organization. Most significanty, the organization lacks the modeling capability to effectively argue its case for rules governing patient movement that would facilitate better service.

Fleet size tends to be fixed (or diminishing) in the short run, for a number of reasons:

1) Aircraft purchases require long lead times, because of the costs involved, extensive specification, selection and approval actions, and lengthy construction and outfitting.

2) Other organizations handle most of these procurement actions. Most importantly, final purchase authority lies outside the organization.

3) Only when the useful life (typically 30 years or more) is effectively exhausted, the technology so outmoded, or spare parts, fuel or maintenance support so expensive or difficult to obtain is an aircraft replaced, particularly non-combat aircraft.

4) Single or small lot purchases of aircraft requiring unique or special outfitting can be prohibitively expensive.

5) Any purchase requires convincing arguments and a demonstrated need over other requirements, both within the Airlift Command and among all other Air Force commands, to be able to compete for limited acquisition dollars.

Only two aircraft models have been used exclusively for aeromedical transportation. The current all-jet fleet replaced an aging and technologically obsolete piston-engined predecessor shortly before the Air Force retired all non-jet aircraft for lack of sufficient fuel supplies.

The current fleet of 21 C-9A's was purchased between 1967 and 1971. No other aircraft have been purchased since then. In fact, decision makers at higher Air Force levels reassigned two aircraft to non-medical uses. A third aircraft crashed and was totally destroyed. Six aircraft are dedicated to aeromedical use in Europe and in the Far East. Until the current fleet approaches the end of its serviceable life at least ten years from now, the likelihood of fleet expansion or replacement is very low.

Temporarily expanding capacity by using the C-141 for high density trunk routes interconnecting major regional facilities is being considered. However, the C-141 is two to three times more expensive to operate, per flying hour, than the C-9. The C-141 carries patients less comfortably and lacks much of the specialized medical equipment available on the C-9. Leasing commercial aircraft is not particularly feasible, since they lack necessary electrical, oxygen and vacuum systems, and would have to be modified to carry patients on litters. Permanently modifiying commercial airliners and cargo airplanes that could be quickly converted to carry patients in wartime (only) has been proposed, but never done. Small jets used primarily to transport senior officers have air ambulance capability, but are not used routinely for that purpose.

An extensive set of policies, directives, regulations and standard operating procedures govern system operation. The organization participates in formulating and defending budget proposals, and in seeking changes to allocations of approved budgets, but the annual spending limit imposed by the US Congress is binding by law. The Congress monitors adherance to such limits, and on ocassion directs its investigative arm, the Government Accounting Office, to examine the organizaton's use of its resources. [JONE82]

Federal Aviation Administration (FAA) regulations, or approved (and usually more restrictive) Air Force substitutes govern flight operations, crew qualification, and aircraft airworthiness certification. DOD proscribes the authorized uses of the airplanes, particularly patient movement eligibility criteria and system performance standards. FAA rules are amplified by Air Force flight rules and operational procedures. These in turn are expanded by the Military Airlift Command to establish mission management, crew training and qualification, flight planning, and patient service requirements, ranging from the very broad to the very specific. Particularly since the system's passengers are patients, these rules and regulations are far more restrictive than the technological limitations of the aircraft require.

The following are some of the more significant rules governing aeromedical operations: [WOOD78,JONE82]

1) The maximum crew duty day, which begins when the crew begins preflight planning two hours before their first flight and ends when the last segment is completed, is sixteen hours.

2) The maximum number of stops on a route is eight for 14-hour and five for 16-hour crew duty day, less one stop if bad weather is encountered.

3) The maximum aircraft range is 2100 nautical miles, subject to restrictions on required reserve fuel at the destination and winds.

4) The maximum speed of the aircraft is 500 knots; actual elapsed time between two airfields also includes departure, arrival, climbs, descents and other maneuvering at less than maximum speec so effective speed is considerably less.

5) The current standards for fueling, ground servicing, and patient handling are 50 minutes when refueling is required, and 20 minutes when not required.

6) Under routine service rules, patients must be picked up within 72 hours of movement validation, and delivered within 72 hours after pickup. Priority service reduces both limits to 24 hours. Urgent cases require immediate dispatching or rerouting of an aircraft.

7) Although not a strict rule, patients should not have to make more than two overnight stops. There is no stated maximum individual patient time on-board an aircraft, but it should be minimized. Priority, urgent and special routine cases may require direct service.

8) Maintenance contracts specify that six aircraft per day are available for missions, one for local training on standby for urgent requirements, one to replace any of the six that develop problems, and three in scheduled maintenance.

Many of these restrictions are based on human capacities and limitations. Man is biologically designed to travel at four miles per hour, not four hundred. Flying disturbs the biological (circadian) rythmn of patients and crews, which necessitates limits on the length of flights or sequences of flights and on the interval between flights required for recovery. Flying is inherently fatiguing, both physiologically and mentally. Cabin air is dehydrated to a typical humidity level of 1-2 per cent, causing the body to lose fluids. Crew members spend considerable time at cabin altitudes several thousand feet higher than that they are accustomed to on the ground.

Several policies have been imposed to insure that patients are moved with minimum disturbance. This apparently accounts for the desire to minimize the number of en-route and overnight stops. Reducing stress on both patients and crews is also emphasized. Stress has several environmental sources. Aircraft malfunctions can cause problems that range from the merely annoying to critical and even life-threatening. Flight crews must contend with such environmental hazards as airfield limitations (runway length, aids to navigation, etc.), terrain, adverse weather, restrictions to visibility, and other aircraft, as well as in-flight patient emergencies, routing changes and other exigencies requiring immediate attention and action.

Aircraft operations are limited by the skill and experience of the crews that fly them. To minimize skill differences and ensure that desired levels of performance can be met in both peacetime and wartime, standardized training, evaluation, and recurring accomplishment programs are followed. In fact, the number of crew members to be trained is the principal determinant of the annual program. In that sense, at least from a budget standpoint, patient service is a by-product of the training program. Because program size has not changed in five years, we presume that indicates it is sufficient for training purposes, and is therefore not a significant limitation. In terms of patient service, increased demand suggests otherwise, but no major increases have been obtained. Other human factors limit aircraft operations, including morale (e.g., not scheduling crews for extended periods away from home), and creating daytime schedules to avoid more hazardous night operations to the maximum extent possible.

Unlike trucks or buses, aircraft are not constrained to a fixed road grid. As a result, planning aircraft routes in three-dimensional space is much more difficult. Each flight segment consists of departure, enroute and arrival phases. Each phase in turn involves a number of possibilities. Departures can be by optional or mandatory FAA or Air Force published departures, or by pilot-requested routings. Enroute segments may be direct (great circle) courses, via the FAA airway structure, or a combination of these. Arrivals can be published Standard Terminal Arrival Routes, FAA-directed, or pilot-requested routings. In some instances, particularly between major airports, entire routes may be dictated by the FAA.

To further complicate the difficult combinatorial route planning problem, optimum route selection further unfolds into problems involving aircraft performance, weather conditions, the status of navigation and communications aids, and airfield characteristics. Aircraft performance is a function primarily of climatic conditions, operating technique, and aircraft characteristics. Although two of these input factors cannot be controlled, performance can be determined, in that engineers have created both manual and automated models of airplane behavior. Operating technique refers to the selection of airpseeds, climb and descent profiles, and engine power settings. The most important aircraft characteristics for flight planning purposes are weight (that varies with the number of people and amount of fuel carried), and aircraft performance.

Climatic conditions are particularly significant in route selection. Adverse conditions (storms, turbulence and icing) can restrict or preclude operations into and out of airports and through enroute areas. Temperature and atmospheric pressure affect engine performance. Winds alter aircraft speed over the ground. For safety and patient considerations, if the weather at three stops is below a specified visibility minimum, the maximum number of stops on a route is reduced from eight to seven. [WOOD79]

Airfield characteristics include operating policies, hazards and restrictions to flight and ground operations, fuel availability, navigation and communication facilities status, and other factors that can inhibit or prohibit airfield use by the wing's aircraft. The C-9A requires a minimum runway length, depending upon aircraft takeoff or landing weight, weather conditions, etc., a minimum runway load bearing capacity, and taxiways and parking areas with sufficient capacity and clearance for maneuvering. There are currently some 13 DOD treatment facilities in the US the system serves that lack airfields meeting minimum standards for safe operations. [DEPA83] Other airfields only allow operations during daylight, or when weather conditions permit. (The C-9A can operate in very austere airport environments, because it requires no ground support).

Despite the existence of abundant data on environmental factors that matter, perhaps the most significant reality is the lack of an adequate environmental information system to capture and manipulate it. Although automated systems (providing current weather conditions and forecasts; navigation, communication, and airfield status notices; and aircraft performance data) do exist, (1) they are not integrated, and (2), planners do not have direct access to them. Instead, planners and crews rely on manual methods to plan and operate missions. This shortcoming cannot easily be remedied within the organization.

2.2.5 Resources and Components. Resources are the means through which the system attains its goals, objectives and ideals. [CHUR69] Resources are within the system in the sense that they can be controlled; they can be consumed, changed, conserved, and most importantly, used to the system's advantage, at the discretion of system decision makers. One principal task of a planning system is to find ways to reduce resource use without degrading client service or to improve service without consuming more resources.

Figure 2.7 attempts to capture the rather convoluted and disaggregated DOD organizational structure to which aeromedical resources are allocated. The aeromedical system uses five major types of tangible resources: people (flight and medical crews, management, maintenance, and support personnel); aeromedical staging facilities; aircraft; physical plant; and budget. In addition, the organization has certain intangible resources at its disposal, including information and knowledge gained from over forty years of experience. The purpose here is not to produce a concise inventory of every asset, but to indicate the ownership, roles and importance of each type.

At the top of the structure, the Assistant Secretary of Defense for Health Affairs is principally responsible for planning and policy, which we will discuss in the next

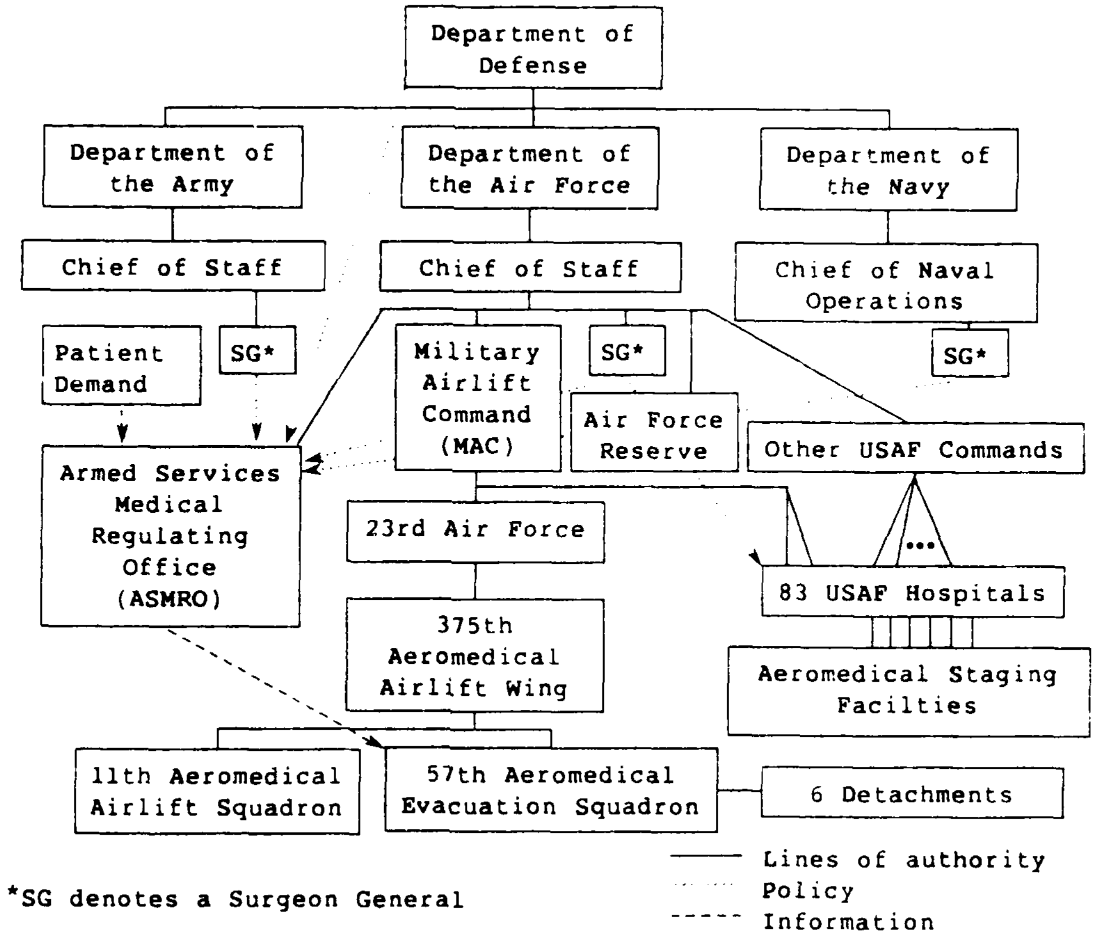

Department of Defense
Department of the Army
Department of the Air Force
Department of the Navy
Chief of Staff
Chief of Staff
Chief of Naval Operations
Patient Demand
SG*
Military Airlift Command (MAC)
SG*
SG*
Air Force Reserve
Other USAF Commands
Armed Services Medical Regulating Office (ASMRO)
23rd Air Force
83 USAF Hospitals
375th Aeromedical Airlift Wing
Aeromedical Staging Facilties
11th Aeromedical Airlift Squadron
57th Aeromedical Evacuation Squadron
6 Detachments
Lines of authority
Policy
Information
*SG denotes a Surgeon General

Figure 2.7. Aeromedical Organizations.

section. This level has overall authority and responsibility for resource use, and for presenting annual resource needs to Congress through the budget process. DOD guidance constrains and directs aeromedical resource use through the policies DOD issues. Not shown are other federal agencies (e.g., the Veterans Administration and the US Public Health Service) whose policies directly affect aeromedical transportation and which both consume and reimburse DOD for aeromedical services used.

The next two levels in the hierarchy provide the overall management and direction of the separate services' medical programs. At this level, policy and guidance become rules and regulations on resource use, and the resource allocations are made from Congressional authorizations and appropriations.

The first point where any semblance of integration takes place is at ASMRO. As the exclusive patient movement regulator, ASMRO has a significant impact on patient service. ASMRO manages one resource, the patient movement request data base, capturing and passing data to the 57th Squadron, which controls the actual transfer process. ASMRO has a staff of 15, a small building at Scott, the computer-based patient data system, and little else beyond office equipment for its critical role. ASMRO operating costs are not reflected in the aeromedical budget.

The remainder of the organizational entities involved in aeromedical transportation fall into two groups. Within DOD, the Airlift Command is responsible for virtually all air transportation, including aeromedical evacuation. Seven members of the MAC Surgeon General's staff are responsible for oversight of the aeromedical function, budget formulation and submission, and the formulation of plans and operating rules. 23rd Air Force is a new middle echelon command and administrative entity to which the

375th Wing was recently assigned. No one at 23rd Air Force is assigned exclusively to aeromedical operations.

The 375th Wing is the principal operational organization, and comes close to being a component by Churchman's definition. [CHUR69] Under the DOD major force program concept, wings are the smallest units with complete mission capability and responsibility. Three groups within the wing, Operations, Aeromedical Services, and Maintenance, totaling 297 people, are responsible for flight scheduling and operations, patient service, and aircraft upkeep respectively. The wing currently assigns 26 officers, 26 enlisted members, and 26 civilians to the first two groups. The Maintenance function, which has 6 officers, 154 enlisted airmen, and 59 civilians assigned, also furnishes 23 flight mechanics to provide one mechanic per mission for enroute aircraft servicing and minor repairs.

Because the wing manages six distinctly different missions (another is an executive jet transportation service for senior military officers, DOD civilian employees, and occasionally, Members of Congress), these numbers reflect those dedicated primarily to the aeromedical transportation program. The numbers also include a pro rata apportionment of people employed in other functions at Scott Air Force Base, such as civil engineering, medical services, that support all six mission areas.

The two squadrons are assigned medical and pilot crews (2 nurses and 3 medical technicians in each of 23 medical and 23 2-person pilot crews) for aeromedical operations, in a ratio of 2.0 medical pilot crews per airplane owned (excluding 6 assigned overseas and two temporarily assigned to VIP transportation). In addition, 3616 Air National Guard and Air Force Reserve personnel provide an additional 598 medical and 18 pilot crews; and 31 liason, 10 evacuation control center, and 15 mobile staging facility teams for command and control functions, patient handling at ground stops, and administrative tasks respectively. They are primarily for wartime system expansion, consituting 92 per cent of all flight medical crews, for example, but they rehearse their roles in normal peacetime operations and training exercises. Because reserve forces are assigned to a major defense program different from their active duty counterparts, the Air Force Reserve reimburses the aeromedical program for their training. This means that the cost of reserve pilot crews flying patient service missions are not reflected as an aeromedical program expense.

Of 83 Air Force hospitals and clinics in the US, five have staging facilities for enroute patient care. We consider the staging facilities to be components, since one of the measures of performance, the number of overnight patient delays, is directly related to their functioning, which we describe later in more detail. The largest, at

Travis Air Force Base, employs 61, 10 officers (primarily nurses), 46 enlisted medical technicians and administrators, and 5 civilians. All five employ a total of 140. Administratively, they are assigned to the local hospital, only two of which belong to MAC. Although exclusively dedicated to aeromedical service, the costs of the staging facilities are paid by their parent hospitals and not out of the aeromedical program budget. In Fiscal Year 1977 the estimated cost was $3,070,000. [DEPA78b]

Detachments are small units of the 57th Squadron located at each staging facility, at Buckley Air National Guard Base, and at McGuire Air Force Base, NJ (where patients embark for overseas destinations). Five officers and 35 enlisted personnel provide administrative, liason, and patient manifesting and baggage handling support.

As a general comment, it should be clear that the organizational alignments of the people involved in aeromedical transportation greatly muddle the question of who does what. One of the consequences is a substantial understatement of the true costs of the system, particularly opportunity costs. The US Congress has made repeated attempts to remedy this situation by consolidating health care into a single agency, as it did in 1947 to form DOD itself from the separate service branches, and has done to create the Defense Logistics and Communications Agencies. However, repeated attempts to form a single defense health organization have all failed, including the latest attempt by the 98th Congress. [HARR84]

The six aeromedical staging facilities (Figure 2.8) accomodate patients during overnight stays while enroute. Staging facilities have the following characteristics:

1) They are minimum care facilities; i.e., they do not provide treatment.

2) They are located at points that function as origins and destinations for patient movements, intra- and interregional transfer points, and overnight stopovers for crews and aircraft.

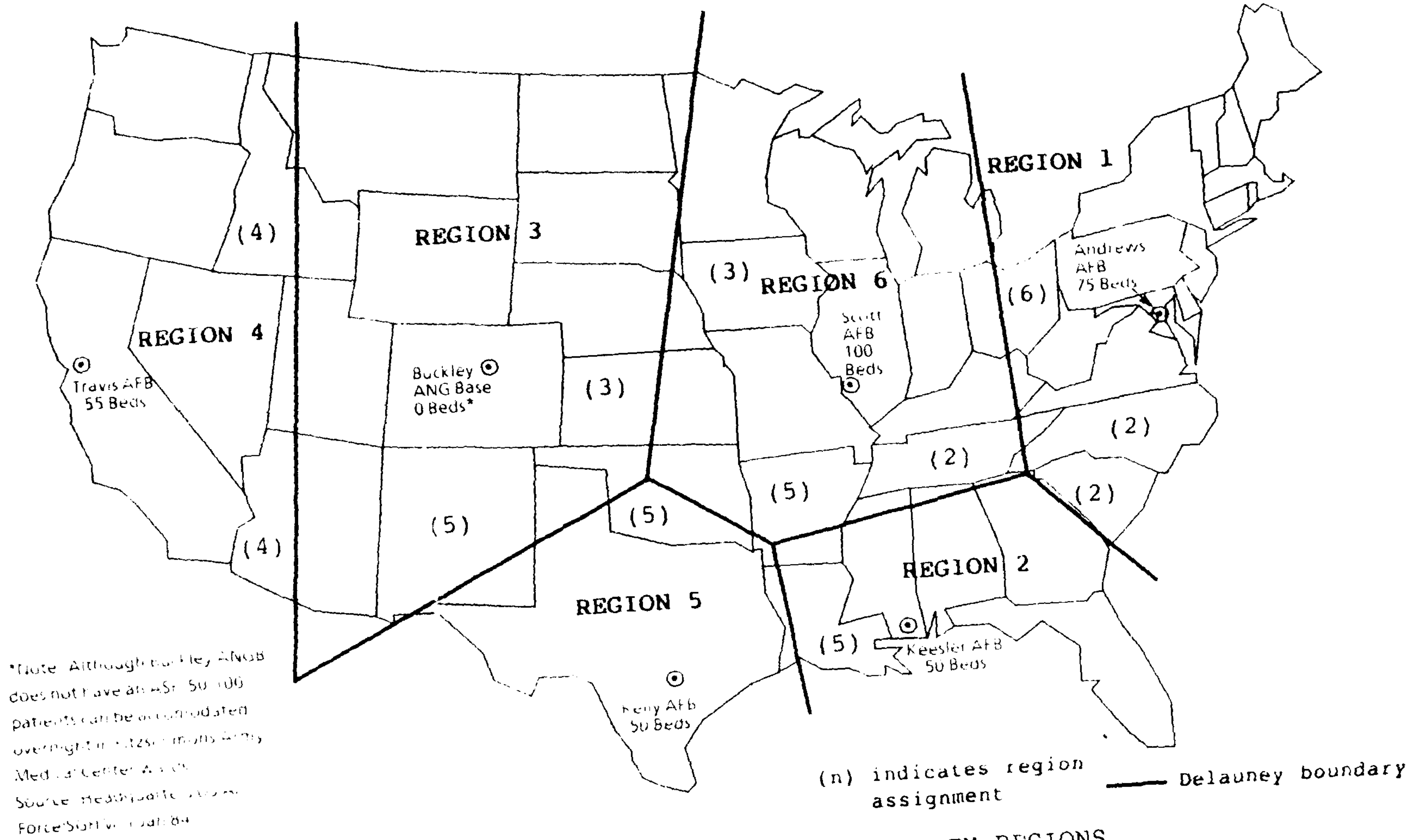
3) Two (at Travis and Andrews Air Force Bases) receive patients from the international system.

4) Geometrically, they the central places in a Dirichlet regionalization of the 48 US states .

5) Capacities shown are total beds available, unrestricted by patient category or condition.

6) Wartime reserve mobilization would expand the capacity of these facilities and activate four more facilities at other Airlift Command bases.

The first point is self-explanatory. Bases with staging facilities serve more patient origins and destinations than any others. 2069 patients, 28.26 per cent of the 7321 transferred during the period we studied, originated at one of the six bases, and 4407, 60.2 per cent of all patients moved, were destined for hospitals served by these bases. As major patient sources and sinks, they are central places in the flow network. All patients coming from outside the US and from the Carribean entered the domestic system at Travis or Andrews Air Force Bases.



*Note: Although Buckley ANGB does not have an ASF 50-100 patients can be accommodated overnight in Fitzsimmons Army Medical Center A [illegible]
Source: Headquarters [illegible]
Force Surgeon [illegible] Jan 84

Figure 2.8. DOD DOMESTIC AEROMEDICAL SYSTEM REGIONS

Figure 2.8 shows the regional assignments of the 48 states. The dark lines are perpendicular bisectors between adjacent ASF pairs. 36 states are assigned to the geographically closest ASF. Six were subdivided by these geometric boundaries and were assigned as shown. Six are assigned to more distant facilities. This might be partially explained by the transfer several years ago of the region 2 ASF from Montgomery, AL, to Biloxi, MS. With Montgomery as the ASF, Louisiana, Tennessee, and South Carolina would have been "properly" assigned, leaving only three not assigned to their closest centers. No one in the organization could find any historical reference for the original assignments. DOD uses a nine-region system to group military medical facilities, but that alignment does not appear to have any substantive aeromedical purpose.

In 1982, the average number of beds occupied at each facility was 18.0 at Kelly AFB, 48.9 at Scott AFB, 26.5 at Andrews AFB, 18.6 at Travis AFB, 16.4 at Keesler AFB, and 0.3 at Buckley ANGB. The average length of stay at each ASF was 2.11, 1.28, .64, .94, 1.34, and .9 days respectively. In 1982, on some weekdays, the ASFs were full. [DEPA83] As we mentioned in our discussion of system purposes, both the capacity problem and the desire not to impose overnight delays on patients have lead to management initiatives to reduce ASF use.

The eleven C-9A domestic system aircraft provide the means to transport patients quickly up to 2000 nautical miles non-stop. They were permanently modified at the time of purchase with special oxygen, electrical, vacuum and other medical systems and equipment, boarding ramps, seats for patients in full orthopedic body casts, and accomodations for litters (stretchers). Each aircraft can carry any combination of 40 ambulatory and litter patients. While no major modifications (e.g., improved engines and electronic fuel management equipment) to increase operating efficiency have been made, medical equipment is continually upgraded. On-board communications equipment permits continuous in-flight coordination with ground facilities, and dynamic rerouting as conditions change.

All routine aircraft maintenance service, including parts and supplies, are provided by a civilian contractor at the home station at Scott Air Force Base, IL. In addition to the aircraft, the wing keeps specialized tools and ground support equipment at the central base. Widespread commercial DC-9 use and the proximity of Scott to Lambert-St. Louis International Airport where the DC-9 manufacturer is headquartered and a major air carrier maintains a DC-9 fleet make large spare parts stocks unnecessary. Because of small fleet size, new pilots are initially trained by commercial contractors, who also provide flight simulator facilities for recurring training.

Under the terms of the maintenance contract, the wing has six to seven aircraft available for daily missions, one for both training and emergency flights, three undergoing minor maintenance (with one available to replace any of the mission aircraft in the event of unscheduled maintenance problems), and one aircraft undergoing major maintenance (lasting 45 days) at another contractor's facility. Mission aircraft represent the means to fly approximately 42 to 48 route segments each day, seven days a week. Fewer than six missions are flown on weekends, patient movement demand permitting, but not more than seven even when demand increases. Besides providing patient transportation, missions provide regular and frequent flying experience to maintain both active duty and reserve crew currency.

The physical plant is relatively modest, because the wing does not maintain permanent facilities away from the central base, except for the six detachments. The ASF's are maintained by the medical facilities at which the ASF's are located. At Scott AFB, the wing maintains all base facilities, although most are used for other purposes. A headquarters building, two squadron buildings, one aircraft hangar, and flightline facilities for C-9 maintenance and aircrew training are the extent of its physical plant. The costs of acquiring and maintaining these assets are accounted for, and a proportionate share of other base operating costs not directly part of aeromedical transportation operations are accounted for as non-direct operating costs.

Each year, as we noted in the last section, the wing receives a certain budget authorization and appropriation as part of the annual DOD budgeting process to pay for the various resources we have just described. Through the Programming-Planning-Budgeting System (PPBS), the wing requests and receives a budget (more correctly, authority to obligate the US Treasury to pay amounts up to that share) to fund aeromedical operations. There are several significant features of the budget that should be noted.

The US Congress and the Department of Defense use two distinctly different formats. The DOD PPBS format is organized into ten major defense missions, including aeromedical transportation. Congress utilizes the traditional line-item format, with categories for personnel, military construction, operations and maintenance, without specifying how they are to be used in mission terms. Costs can be determined by referring to either format, although each offers better and worse features than the other.

Each year the budget process maps resource requests into the program budget format, then into line items for Congressional authorization and appropriation, and then back into the ten programs. As part of that last process, the wing's obligational authority for various kinds of resources is pooled into the maximum number of flying hours[21] the wing can fly to accomplish its mission. Programmed total direct operating costs for 25806 flying hours in the 1985 fiscal year budget are as follows:

Cost Category (in $1,000s)	Fixed	Variable	Total
Aircraft operations			
Depot Maintenance	$13295		
Supplies and equipment	880		
Engineering services	1048		
Jet fuel[1]		$24516	
Crew Per Diem		1001	
	$15359	$25517	
Aircraft operations subtotal			$40876
Personnel			
Civilian pay	$2563		
Foreign workers	162		
	$2725		
Personnel subtotal			$2725
Administration			
Office expenses	$16		
Travel (non-crew related)	651		
Utilities	577		
Communications	77		
Real property maintenance	221		
Base operating costs	668		
	$2210		
Administration subtotal			$2210
Total Operating Cost (excluding pay)			$45811
Less reimbursements for non-medical passengers carried and for reserve and guard salaries			-1672
Total Direct Operating Costs			$44139

NOTE: 1. Jet fuel is charged at one standard price at all Air Force installations in the continental United States.[21]
Source: Headquarters MAC/SGM, 19 April 84.

Each flying hour, then, "costs" about $1700; in fact, that amount plus the direct expenses of military pay and allowances is the charge to any authorized users who must reimburse the wing. It is interesting to note that these costs do not include other charges for aeromedical care costs ($105), factored amounts to recover military pay and allowances ($770), other personnel costs ($120), and military moving costs ($50); which add an additional $1045 to the charges made to other non-DOD, federal users. To other users not eligible for US government rates, the additional costs of retirement entitlements ($214), administrative expenses ($23), and a 4% asset use charge (in lieu of aircraft depreciation) ($115) are added, for a total of nearly $3100, plus any charges for special requirements or other abnormal costs. [DEPA84,p.84] (ASF costs are not included.)

There are a number of salient observations that we should make regarding system resource and component costs.

1) Actual costs are considerably understated. Factor costs are not all accounted for, even in the non-government charge. Notably absent are the costs of personnel assigned to 23rd Air Force, MAC, ASMRO, and the two DOD levels. The Air Force Reserve provides reserve salaries, and the parent hospitals pay their staging facilities' operating costs.

2) DOD uniform fuel pricing discourages economic operation. The policy to carry full fuel loads maximizes flexibility but hinders conservation, as does the lack of flight planning automation and on-board fuel management equipment. The airlines, for whom fuel is the largest variable operating cost, track fuel use closely, by plane and even by pilots, to detect patterns of uneconomical practices.

3) Both the referring hospital and the patient have strong incentives to choose air transportation to another facility over local civilian treatment. Referring hospitals must pay for the full cost of local treatment out of limited budgets, and all but active duty patients must pay for some portion of it. The air transportation system and the destination military hospital absorb all costs, making transfers essentially free to referror and patient. Comparing costs is difficult, because of understated flight costs, the flight cost allocation problem, and the lack of accurate care cost data at DOD hospitals. [FRAG82]

4) There is a strong incentive to maintain slack to absorb the unanticipated.

5) The principal control device emphasizes inputs, not outputs.

6) The rationale behind allocating resources to servers rather than users is to produce training outcomes; user service is a fortuitous by-product. The allocation mechanism has no pricing cues; no competitive alternatives are available or permitted to provide market prices, and no internal transfer pricing via working capital funding is used. Performance measures relate resource inputs to service outputs, expressed in highly aggregated, service-oriented terms, e.g., numbers of patients moved, not training outputs.

Though difficult to quantify, the aeromedical organization also has significant intangible resources. In their 40 years of moving large numbers of patients, their accumulated experience and knowledge undoubtedly enhances service quality and system capability. The data they have collected provides an historic record of resource use and patient service, though it is not currently used to best advantage to improve resource allocation and utilization. Finally, one has to be impressed by the spirit, enthusiasm and dedication of the people who are convinced that they ought to be doing what they are.

2.2.6 Planners. One of the most fundamental questions in designing planning systems is, Who should be the planners? One approach would be to use the mapping approach we employed earlier for clients to answer the related question of who they currently are. By identifying current planners and their qualifications, and determining how and by what right or authority they perform their planning functions, we should gain some insights into who they ought to be. As seems a familiar pattern, however, that determination ultimately requires choice, which unfolds into answering other difficult questions of how we measure or judge the competence of planners, their performance, when and what recourse or compensation would be appropriate should their plans fail, and the efficacy of the approaches and methods they employ. Apparently no one has proposed any defensible criteria for making such judgements. [CHUR79]

For that reason it seemed most reasonable to restrict our attention to planners and planning decisions directly involved in patient treatment, perhaps the most significant of the transportation system's purposes. The ability to care for patients is function of the three major treatment planning variables: capability (which services will be provided), capacity (to how many), and entitlement (who is eligible to be served). Transportation needs are directly related to imbalances in the distribution of capability and capacity among facilities and to limitations imposed by entitlement rules. We might have simply assumed these conditions to be dictated by an environment in which scarce resources and other unalterable factors create these conditions. But, in examining the roles of the the planning groups listed in our map in Table 2.6 we found that to a significant extent these conditions resulted directly from planning failures caused by participants in the aeromedical transportation planning process.

Congress legislates health care entitlements, presumably by considering the need to maintain the health of those providing the nation's defense and to provide an adequate health benefit as compensation for that service. The entitlement rules create three classes of patient service:

1) Active duty soldiers receive all in-patient and out-patient care without cost.

2) Active duty dependents receive unlimited in-patient services under CHAMPUS at civilian hospitals, or *on a space available basis* at military facilities, for a nominal daily charge. Out-patient care is free at military clinics; dependents repay CHAMPUS for a deductible and a co-insurance portion of charges at non-military clinics. (In 1979, the three charges were $4.65 per day, $50 per person and $100 per family, and 20 per cent of actual costs, with no upper limit). [DRMS79]

3) Retirees and their dependents have no guarantee of service at a military facility even if there is space available for them, and they must pay the nominal daily charge for in-patient care. Under CHAMPUS, they must pay the deductible for out-patient care, and 25 per cent of both in-patient and out-patient charges. And, at age 65 they become ineligible for both CHAMPUS and Medicare.

TABLE 2.6

DOD AEROMEDICAL TRANSPORTATION PLANNING MAP

Phase	Agency	Inputs	Constraints	Output
Legislation	Congress	DOD Budget Requests	Political Economic Societal	Defense Missions Entitlement Rules Budget
Policy Formulation	DOD	Budget Defense Missions	Entitlement Rules	DOD Policy DOD Health Missions Budget Shares
Policy Enactment	DOD Health Council	DOD Health Missions	Fiscal Legal DOD Policy Guidance	DOD Health Policies Service Criteria Regulation Policy Regional Boundaries
Program Formulation	Services	Budget Shares	DOD Health Policies	Service Health Programs Facility budgets Operating Rules
Program Execution	Hospital	Health Programs Budget	Operating Rules	Treatment Programs
Regulation	ASMRO	Movement Requests	Regulation Policy Regional Boundaries Facility Needs and Capacity	Mission Parameters: Origins Destinations Categories Special Needs
Movement	375th Wing	Mission Parameters Resources	Service Criteria Patient Needs Operating Environment	Movement Schedule

In its examination of the DOD health services system, the Defense Resource Management study concluded that

> At one time the military services were generally considered to offer the best medical benefit available. In recent years, though, military personnel, retirees, and dependents appear to be increasingly dissatisfied with their health care benefit. Unavailability of services, long queues, attitudes of providers, administrative mixups, and excessive costs of CHAMPUS, are the most strident and most frequently heard complaints. ... While the CHAMPUS program reimburses those who are denied in-house care for much of the cost of civilian care, the CHAMPUS program can be distinctly inferior to in-house care in financial protection, covered services, convenience, continuity, and quality of care. [DRMS79,p.94]

More significantly, the study concluded that

> The bifurcation of the health care system into CHAMPUS and direct care allows patients to switch or be switched from one branch of the system to the other, potentially resulting in harmful or demoralizing discontinuities of care. Current regulations prohibit military physicians from making direct referrals. Hence, if a local military clinic or hospital lacks the necessary services to treat a beneficiary's illness, the military physician must either refer the patient to a military hospital that offers the required service (even if the hospital is far away from the patient's home), allow the patient to find his own care in the local community without referral advice, or attempt to persuade his or her commanding officer to have the clinic or hospital pay for the civilian referral. In the latter case, there is little advantage to the commanding officer to pay for the care out of his or her own funds when CHAMPUS will pay. In the other cases, the patient suffers. The patient either must be uprooted from home for treatment or must find the right phsician or hospital, perhaps in an unfamiliar locale to which the patient happens to be assigned. If the patient chooses to seek local care, the patient may never be referred back to his or her primary military physician, further inhibiting continuity. [DRMS79,p.106]

As another direct result, many beneficiaries receive no care. [DRMS79] However unintentional, current entitlement planning has created several very negative treatment conditions, including uprooting patients from their homes and sending them to distant facilities, forcing them to choose transportation as the lesser among many evils that include no treatment at all, and saddling them and the taxpayer with unnecessary travel costs. These circumstances present the transportation planner with substantially more difficult problems than simply resolving supply-demand medical service imbalances.

These problems would be mitigated, in all or at least partly, by adequate capability and capacity planning. But, there is substantial evidence of significant shortcomings here also. Defense planning includes allocating medical resources between the two major health missions and distributing resources in sufficient quantities and to appropriate locations that meet both peacetime and wartime planning demands, which as we discussed earlier inherently conflict with each other. As studies of defense planning and budgeting by Crecine [CREC71] and Wildavsky [WILD75] have concluded, however, even under the appearance of rationality afforded by PPBS, the process is more one of reacting to the political and economic environment, "muddling through" [LIND59] rather than comprehensively evaluating and selecting the best planning alternatives.

Several specific deficiencies in capacity and capability planning have been identified that are related to the organizational and institutional factors we described earlier. At the DOD level, responsibility for the two principal health missions must be assigned to the services. DOD clearly and unambiguously assigns readiness objectives, missions, and planning responsibilities to the service Medical Departments through its directives and regulations. But as the Defense Resource Managment Study found, "Nowhere in service regulations was the DRMS able to find explicit assignment of the benefit mission to health care managers." [DRMS79] As a direct result,

> The absence of an explicit health benefit mission leads to an unnecessarily convoluted logic train in the justification of resources needed to accomplish the health benefit mission. [DRMS79,p.82]

In addition to creating problems in providing adequate resources for health benefit treatment, this situation induces bias toward the readiness mission in allocating, distributing, and using medical resources:

> Health care tends to be viewed by the managers of the system not as a guaranteed benefit at some specified level but as a serendipitous by-product of a health care establishment that exists to maintain the health of the active duty force and to provide wartime support. [DRMS79,p.94]

Other studies have noted contributing problems. An OMB/DOD/HEW Study in 1975 concluded that "the M[ilitary] H[ealth] S[ervices] S[ystem] is handicapped by lack of adequate population, workload, and cost data and comparable information systems for the Military Departments." [DRMS79,p.112] Developing a unified, standard cost accounting system could greatly improve the collection of actual direct care costs necessary for comparative analyses of direct and civilian treatments, but the prototype is still incomplete. [FRAG82] Efforts to determine the population-at-risk, actual numbers of eligible beneficiaries, the service cachement areas of military facilities, and actual and forecasted workloads of all MHSS facilities are underway, but not yet completed. [DOYL82,STCL82] Until these problems are resolved, adequate capability and capacity planning is impossible.

Decentralized and separate medical services are a longstanding DOD tradition. They also seem to be major contributors to the situation where only a small fraction of all patients (fewer than 22 per cent, on average, in calendar years 1981, 1982, and 1983 [HARR84]) are moved to the facilities of another service. With the possible exception of flight and submarine-related clinical specialties, health care is one of the most uniform of all missions across the services. However, with the emphasis placed on maintaining the integrity of medical units for deployment purposes, attempts to consolidate DOD health care have all failed, at some unknown cost in underutilizing available capabilities and capacities.

Another instance of lost utilization opportunities occurs in transfer regulation, because of the traditional practice of not allowing regulators to intervene in the determination of patient destinations. Siverd [SIVE78] predicted in his study of regulation policy and its impact on patient service that considerable improvement in patient transfers could be realized if the regulation process were interactive among the facilities, doctors and ASMRO. In the next section, we will illustrate how changing the planning criteria for destination selection during the regulation process might improve transportation service.

For sake of discussion, Table 2.7 groups the various planning agencies according to our understanding of their principal planning concerns and roles, and relates them through a continuum based on client welfare. Roles are not exclusive and disconnected; rather, the overlapping and conflicting concerns of serving all versus serving each individual patient must be balanced, particularly by planners in the middle group.

This scheme also introduces the patient as planner, a role that may seem somewhat curious, and is often ignored, since most system participants presume that the "expert" in patient treatment is the doctor. But, his world view is that of the any specialist; the patient is ill, so he must be treated. In seeing that the patient receives the best

TABLE 2.7

COLLECTIVE VERSUS INDIVIDUAL PATIENT PLANNING CONCERNS

Health Program Planners	Movement Planners	Treatment Planners
Congress DOD DOD Health Council Surgeons General of the 3 services	ASMRO 375th Wing	Hospitals Doctors Patients
All Patients	↔	Individual Patients

and most appropriate treatment, the doctor fully discharges his planning duties.

However, to patients, treatment is only one of many personal concerns.

> Health care utilization patterns of individuals have been demonstrated to be the result of a combination of compex factors which include socio-demographic characteristics, expected benefits from seeking care, the availability of sources of care, and the economic and social costs of seeking care. [...] the decision of where to seek health care is as complicated as the decision of whether to seek health care and some of the same factors may be relevant. [STCL82,p.3]

Individual patients would seem the most appropriate "experts" to determine their own treatment objectives and constraints. That would include balancing the costs and other aspects of CHAMPUS care against delays, disruptions and inconveniences caused by the aeromedical system, to travel to distant facilities when they cannot be provided direct care at military hospitals. An informal 375th Wing survey found that predicted aeromedical transfer time strongly influenced patient choice of treatment plans. [POFF84]. These individual planning decisions strongly affect other planners, particularly those in Table 2.6 who incorporate patient needs into their planning.

In summary, DOD aeromedical transportation planners must resolve planning problems that to a significant extent stem from the lack or failure of other medical planning processes. Congressional entitlement planning forces some patients into the aeromedical system because it creates either inferior or no alternatives, not because transportation is the best alternative to the patient. The health benefit mission is not assigned within the services, whose attitudes and resource allocations are biased toward one mission (readiness) and one beneficiary class. Through its separate service medical organization, DOD tacitly permits practices that emphasize the service's and not the client's interests, by restricting or directing patient transfers along organizational lines. By not permitting ASMRO more direct role in transfer planning, patients may be subjected to unnecessary travel, and opportunities to improve service can be lost. Capacity and capability planning suffer from the lack of adequate cost, population-at-risk, workload, and resource utilization data, resulting in resource allocations that produce service shortages that the transportation system must resolve. Clearly, the planning function ought to be better than it is.

2.2.7 Decision Makers. The transition from a discussion of planners to decision makers is aided by a natural relationship that exists between planners and decision makers: if plans are implemented, the planner in effect is a decision maker. In fact, if we use a mapping approach as we did for planners, we will find that to a great extent the two maps are coextensive. Decision making and planning collide when the time comes for plans to be implemented.

In our definition of the decision maker, we said that it was he or they who ought to coproduce action (with the environment, whose elements are beyond the decision maker's control) through the use of appropriate resources. The decision subsystem takes action by (1) setting the actual levels of planning variables, (2) carrying out the plan at the levels chosen, and then (3) controlling the resulting outcome, monitoring results, comparing them with those desired, adjusting the allocation of resources, and generating new actions, by either suggesting or making changes to plans in the process.

To better understand the identities and roles of aeromedical transportation decision makers, we constructed the "decision map" in Table 2.8, listing all those who influence the system's outcomes. The list's order follows that suggested by the diagram of the major organizational participants (Figure 2.7).

TABLE 2.8

AEROMEDICAL TRANSPORTATION DECISION MAP

Decision Maker	Variables	Decisions
Congress	Economic, social, and political conditions	Defense budget Entitlements and benefits
DOD	Budget levels Defense policies	Mission assignments Budget allocations
Services	Mission assignments Budget allocations	Budget distributions Health programs
ASMRO	Facility capacities and specialities Training program requirements Patient diagnoses	Patient destination Movement validation
MAC	DOD transportation policies and rules Aeromedical budget Readiness goals	Operating policies and plans Qualifications Flying hour allocation
Wing	Flying hour budget Resources available Movement demands	Operating rules System schedule Alterations to daily routing schedules
Evacuation Control Center	Patient condition System schedule	Enroute care Service stipulations
Operations	Operating rules Environmental conditions	Daily routes Command and control Dynamic rescheduling
Medical Facilities	Mission Budget Health programs	Treatment programs Training loads Aeromedical movement or local treatment
Doctors	Patient condition Treatment policies	Diagnosis Treatment alternatives
Patients	Availability and accesibility of care alternatives Perceived benefits and costs of care	Treatment alternative

In improving a system one of the most imprtant tasks is to design functional subsystems, or components, that have the resources and management to carry out one or more complete missions of the whole system. (While this seems to contradict the notion of whole system design, the reality is that we, as designers, cannot adequately create single components to perform all missions of complex systems.) Mission components should be designed so that their performance is directly related to the performance of the whole system. Component performance should reveal how the component operates and if it is operating properly, including its decision making and informational mechanisms.

One of the most common and most serious consequences of breaking down a whole system into parts or subsystems is that the principal concern is problem-solving. The rationale often attributed to an organization's structure is that it represents the organization's view of which problems it considers important (and which are not), an allocation of resources to solve the problem, and task assignments and responsibilities. But, what is commonly missing is the assignment of complete mission responsibility, as we saw in the case of DOD's health benefit mission. The resulting uncoordinated array of organizational pieces addresses only aspects of the health care delivery problem. This fragmentation makes performance measurement extremely difficult.

Presuming that transportation is one defineable DOD health mission, there are two principal decisions involved in carying out that mission: deciding where patients go and how they get there. These decisions are obviously related, but each is sufficiently complex that it would be exceedingly difficult for one component to handle both. The mission of the first component is principally concerned with utilizing facility capacity and capability, supporting training programs, and balancing institutional and individual patient needs. The second component should be concerned primarily with patient needs. While we can find current organizational units that handle portions of these missions, none is given complete responsibility.

The purposes of having smaller, but interconnected components is very different from those of isolating problem areas and improving the efficiencies of individual units tasked with solving those problems. The essence of an efficiency approach is the balance sheet, with its lists of resources available. The basis of any decision is typically cost, measured in terms of using up available resources. As Churchman notes, operating efficiency seems to be an overriding objective of all managers of systems, namely, finding inefficiencies in the form of high costs and eliminating them. [CHUR67] But as Mintzberg argues very persuasively, efficiency approaches have usually failed, and for a number of consistent reasons. [MINT82]

1) Efficiency places great emphasis on calculation, and calculating costs is usually much simpler than calculating benefits. Efficiency naturally tends toward one system of values (economic) and away from others (oriented toward social benefit).

2) In a kind of goal displacement, efficiency measures often become values in themselves, creating a "cult of efficiency". [PFEF78]

3) Efficiency is accounting transaction-oriented; if an event hasn't occurred yet, no cost has been incurred. The approach loses its ability to forecast future events and becomes almost completely historically oriented.

4) Its philosophy emphasizes one best way; inefficiencies (e.g., high cost) are absolute and must be eliminated; solutions are obvious.

Aeromedical decision making is continually critiqued by perhaps the strongest proponents of efficiency, its auditors. To them, unused ASF beds and urgent missions that serve one patient are obviously wasteful and can be eliminated to reduce costs. [DEPA78b] The organization argues otherwise, but with great difficulty, because their response centers on benefits and costs, and the benefits, particularly those that accrue to readiness, are very difficult to measure. Service quality, patient satisfaction, and other qualitative and subjective benefits are likewise difficult to calculate.

A very different decision making approach provides a way to respond, and also attempts to avoid some of the very real dangers posed by the efficiency approach. It is based on the idea of a management information system with some very specific capabilties.

1) The MIS should view all cost as opportunity costs, defined as the "best" alternative forgone when an action is taken, measured in terms of costs and benefits. To estimate this cost, the MIS must contain a conceptual view of the decision maker's model, both its realities (what is feasible) and its values (the benefits involved). [CHUR67]

2) Resources, both in the future and in the past, must be viewed in terms of alternative ways they can be used. The principal MIS function is to make alternative uses explicit and understandable to the decision maker.

3) Decisions on resource use have global, not isolated effects, and must be so evaluated. Inefficiencies are relative and must be balanced. Decisions that lead to overall improvement are often not obvious and non-intuitive; some costs may increase, and "illogical" choices, e.g., longer routes, may yield greater improvement.

Perhaps the greatest danger in an efficiency approach is ignoring benefit concerns. A transactions orientation measures performance not in terms of the legitimacy or importance of meeting client demands, but on the number of demands satisfied. The ASMRO function is a case in point. Nowhere in their measure of performance is client benefit expressed; annual production is a tabulation of patient movements regulated. Health managers do not specifically prevent them from being concerned with the purposes and quality of patient movements, but neither do they rate them on how they are able to incorporate patient welfare in their decisions. The main decision rule is based on efficiency, and apparently, an economic value: move patients to the closest facility with adequate capability. We will illustrate at the end of this section that the resulting opportunity costs may be very high.

A third design isssue is how to map data into information for decision making. Any mapping is based on a particular world view: which data, how data and information are related through values, and what kinds of decision alternatives are sought. The efficiency versus MIS debate illustrates (1) the danger of incorrect or non-appropriate world views and (2) that many views are possible, as Mason pointed out. Differing views promote, rather than eliminate, conflict, but for a good purpose: the debate they induce forces us to recognize important concerns overlooked by a single view. Correctly designed, a decision system based on debating differing views will avoid the efficency practice of relegating benefit concerns to "externalities" and then conveniently forgetting them.

One of the best ways to do this is to create powerful opposites, or dialectics, and embed them in decision systems. We see some excellent possibilities for this in current aeromedical decision making. We observed the needs of individual patients and the needs of all patients being debated by the medical and flight planners during daily planning meetings, although the debate was structured around two different sets of data, one about the patients and the other on airfields and weather and other things important to pilots. In destination planning, no one questioned the view that the nearest destination with capability is the best decision rule for choosing patient destinations; flight planners should be able to debate those choices on the basis of better routing opportunites. While doctors certainly have or should have their individual patient's interests in mind when they select or strongly influence the choice of destinations, this does not suggest that their choices should go unchallenged by others, like the ASMRO regulators, who may see other opportunites from their vantage point of observing many cases with similar diagnoses, or even the patients themselves. Granted, the current system uses the safety afforded by routine and rigidly structured decision making to avoid disastrous results. But, without flexibility, the system cannot find innovative ways to resolve its current problems caused by increasing demand and fixed or declining resources.

What we have found in the problem-oriented (vs. mission oriented) approach is the elimination of overlap and important linkages. Figure 2.9 illustrates one example of the consequences of ASMRO choosing the nearest facility with capability. Each patient origin is shown with two allowable destinations, and assuming the points are arrayed in the Euclidian plane, one destination is closer to each destination. Unless an exception applies, ASMRO must

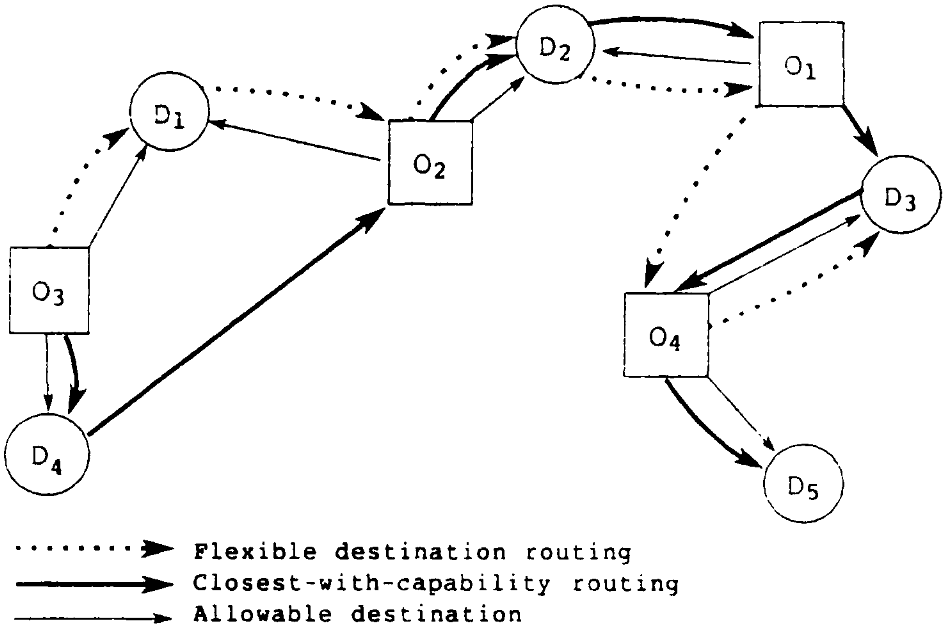


Figure 2.9. Patient regulation decisions.

choose the closer one in each case. Destinations D_2, D_3, D_4, and D_5 would be selected, and the route

O_3 --> D_4 --> O_2 --> D_2 --> O_1--> D_3 --> O_4 --> O_5

would result. If alternatives were allowed, the route

O_3 --> D_1 --> O_2 --> D_2--> O_1 --> O_4 --> D_3

would both reduce the distance traveled, and more importantly, the number of stops. The route

O_3 --> O_2 --> D_1 --> O_4 --> O_1 --> D_3

would eliminate another critical route segment. But, these routing efficiencies must be balanced against the opportunity costs of not sending the patients to the facilities according to the closest distance rule; reducing distance may create adverse effects that outweigh the savings, even though patients seem to benefit.

2.2.8 The System Guarantor. Perhaps the most important aspect of the planner's reality is (or at least should be) whether or not his plans will succeed, which is also (or should be) of great concern to the system's clients. What, if anything, will guarantee success? The optimality theorem of linear programming guarantees that the best solution will be found, if one exists, although we have also talked about how LP solutions depend critically upon the whole system assumption, that the LP model includes nothing irrelevant or erroneous and excludes nothing relevant. LP also requires a certain faith that, if the optimal plan is pursued, the optimal result will be realized, even though the pursuit may extend beyond any previous experience.

We can't overlook an ethical dimension to the optimality question; if we can solve for, pursue and even attain an optimal condition, does that necessarily mean that we should? Careful reflection on, say, an LP maximization of heroin sales or profits suggests that we should not, in some instances. Without trying to be overly dramatic, the point here is that methods such as LP lack the capability to either guarantee success judge what is achieved in terms of client betterment. If we can accept these assumptions LP requires, than the guarantor problem is resolved. But since we have good reasons to think that we cannot, then what provides an acceptable substitute?

There seem to be two troublesome aspects of the guarantor that need to be resolved. The first is that our planning model be internally consistent, that it not contain or lack elements that will create erroneous plans. The second is that the global concerns of the model are correct, that the needs of the clients we seek to serve ought to be served, and that needs satisfaction is equitable among all clients. If those needs ought to be served, then the system ought to survive, which is the most important determination that the systems approach can make.

Some suggest that LP contains a built-in mechanism, sensitivity analysis, to aid in guaranteeing internal consistency and correctness by addressing such questions as what happens when variables are added to and deleted. But all such methods fail for one fundamental reason: they are not able to detect errors in the construction of the model itself. Indeed, all rational models ultimately require external judgements, such as specifying "correct" significance levels in hypothesis testing or "all" the opportunity costs of holding inventories, because the model results are only valid to the extent that the model is valid, which can only be judged externally.

On the second concern, LP does find an optimal value Z, and we can examine its accompanying plan for the collective and individual benefits it provides. But we have no way of measuring how far any one solution may be from the real optimum optimori, or if one is closer than any other, or even if a series of attempts approximates the real optimum. Our plans could lead to disastrous results, but we could not anticipate that outcome, even approximately.

Churchman suggests that

> Some other theory of approximation is needed for all planners, whether or not they employ mathematical programming. After much soul-searching, I've only been able to find one, namely, the theory that a "guarantor" exists, which guarantees that our best efforts, made with every attempt to be as comprehensive and honest as we can, will not contain an error so colossal that our recommendation for action will lead to disaster, and that overall there will be a gradual betterment of the human condition. [CHUR79,p.98]

In other words, the guarantor is not based upon the mathematical properties of the system model, but upon the planner's intentions and efforts. The best we can propose for an aeromedical transportation planning system guarantor, then, is a theory, that the best insurance against disaster is to compel the planner and the decision maker to constantly reexamine and revise their world view, so that they do not commit any errors that can be avoided.

Before we conclude our discussion of the aeromedical system, we will attempt to characterize the movement of patients from historical patient movement experience. Following that, we will then propose ways to improve the system based upon our observations in this section and on the historic functioning of the current system.

2.3 The DOD Aeromedical Transportation Problem. In the first sections of this chapter, we discussed why the DOD Military Health Services System needs to transport patients between facilities. In the second section, we proposed a systems approach to developing a planning system for the aeromedical transportation function. Because plans must be specified in both spatial and temporal terms, we first need to understand those characteristics of patient movement.

There are elements common to all forms of movement phenomena. All forms have a geographical referrent; patient transfers have specific origins and destinations. Movement is purposeful; at one location the supply of a medical service is insufficient to meet demand, while other locations have unused capacity. The spatial interaction between places with supply-demand imbalances can be characterized by a volume of movement (flow) and the expenditure of time, effort and resources (costs). We commonly refer to points as nodes, and differentiate them by function (origin, destination, transshipment point, storage location, etc.). Nodes are linked by routes over which flows occur. Both nodes and routes exhibit structures that we expect to be meshed in an organized way. We call these structures networks, and we expect them to be organized hierarchically, as we would any complex system. We should also be able to observe attempts to economize on time, effort and resources expended in routing patterns, since most creatures, including humans, seem to prefer the shortest, least expensive, or easiest paths when they travel.

In the following sections, we will attempt to characterize patient and aircraft movement in terms of flow and air route networks. We first describe the patients who received service during a ninety day period. We then use both descriptive and explanatory methods to transform patient flow and aircraft movement data into structural and temporal network models. From these representations of the actual system, we will then formulate the model in the final section of the chapter that the remainder of the thesis will address.

2.3.1 Patient Demographics. To determine the nature of patient movements, we examined 14597 individual patient movement data records provided by the US Air Force Military Airlift Command's 375th Aeromedical Airlift Wing, Scott AFB, Illinois, for the ninety-two day period from November 1, 1978 to January 31, 1979. Each record provides details on one patient moved during one mission.[22] If patients transferred to another plane during a mission ('interplaned'), or traveled on more than one day, additional records were generated. The maximum number of records for one complete patient movement was five.

Each record contained the following data: name, rank (or civilian entitlement category), enplaning and deplaning airfields, originating and destination hospitals, patient classification, delays before pickup and while enroute, travel duration, domestic/overseas origin indicator, branch of service, and mission identification. Due to Privacy Act and confidentiality restrictions, information on patient diagnoses and special medical requirements (in cases of burns or other very serious injuries) were deleted from the data supplied to us. (These factors are important in medical flight planning, directly affecting aircraft routing.)[23] Besides the diagnostic information we could not access, there were undoubtedly other unrecorded factors involved in movement decisions, such as humanitarian motives (e.g., moving patients to hospitals near their home towns), and the institutional factors we discussed earlier.

Because the aeromedical data collection system is transaction-oriented, historically recording only actual movements, it does not show movements that were cancelled, or other intermediate decisions that were made and then changed. We therefore do not know how well this data represents true underlying demand, particularly since we could not measure the extent to which knowledge of system schedules influenced the timing of movement requests. The particular time period we chose may not represent other periods we did not select. In the time since our sample was collected, patient movements have increased, at an estimated four to five per cent per year. Therefore, we present the following for descriptive purposes only, to better acquaint the reader with the problem.

Because the patient movement data base did not provide complete information on aircraft routing,[24] we obtained all aeromedical aircraft movement data records for the same period from a different source. This proved particularly fortuitous, since approximately thirty per cent of the patient records contained erroneous patient identification, hospital codes, airport descriptors, and mission data. We assumed that in conflicts between the two data bases, the aircraft movement data were correct; in 5140 aircraft records, we found only two minor errors. In other instances, particularly with multiple mission patient movements and multiple trips by the same individual, we were able to find and correct inconsistencies in patient names, ranks, classifications, personnel categories, and origins and destinations by deduction.

Table 2.9 provides an overview of the patient movement problem during the ninety day period. 1900 "patients" were actually 152 medical and 1748 non-medical attendants. The latter were primarily family members accompanying patients. The attendants group constituted approximately eighteen per cent of all those moved. Their principal impact on the system is using limited aircraft capacity.

TABLE 2.9

SUMMARY OF PATIENTS MOVED BY THE DOD DOMESTIC SYSTEM, 1 NOVEMBER 1978-31 JANUARY 1979

Personnel	Neuropsychiatric			Litter		Ambulatory			Drug	Recovered		Attendants	
Category	I	II	III	I	II	Adult	Infant	Troop	Abuse	Adult	Infant	Medical	Other
Air Force													
Active	31	54	297	52	29	774	0	450	4	282	0	50	266
Retired	2	3	7	37	96	220	0	15	0	162	0	0	91
Army													
Active	70	125	205	81	196	426	0	19	34	81	0	75	142
Retired	0	2	1	39	114	221	0	14	0	128	0	0	61
Navy/Marines													
Active	20	73	411	50	130	280	0	69	124	62	0	21	49
Retired	0	0	1	16	39	75	0	7	0	39	0	0	43
DOD Civilian	0	0	1	2	4	0	0	0	0	0	0	0	2
Dependents													
Active	2	12	19	149	200	469	143	76	0	319	76	1	664
Retired	0	1	6	81	297	488	6	26	0	350	0	1	337
Civilian	0	0	0	1	1	0	0	0	0	1	0	0	6
Revenue-Reimbursable													
VA	0	0	0	1	0	0	0	0	0	0	0	0	3
USPHS	0	2	3	16	12	11	0	0	0	10	2	0	13
Foreign Nationals	0	0	0	0	1	4	0	1	0	0	0	0	2
Other	2	0	0	17	7	6	0	2	0	3	0	4	69
Totals	127	272	951	542	1326	2974	149	679	162	1437	78	152	1748

The remaining 8836 patients can be further subdivided into two groups. 1515 recovered patients (classified in the table as adults and infants) were returned from treatment, typically to their originating medical facility, on an "opportune basis", which means that they were moved by the system as time and capacity permitted. However, they did not receive preference over patients being moved to treatment, and the time they spent awaiting pickup and enroute to their destinations was not measured. Deducting attendant and recovered patient groups from the total leaves 7321 actual patients.

We can further classify the actual patients in ways that relate directly to their handlng. Perhaps the most important is the precedence accorded them by their medical condition. By DOD regulations, patient movements must be approved by the ASMRO office. However, ASMRO only regulates routine, in-patient movements, and simply validates the need to move inpatients at a higher precedence than routine. Priority precedence requires pickup within 24 hours after movement validation, and delivery within 24 hours after pickup. Urgent cases are moved immediately, either by rerouting aircraft, or by dispatching an aircraft

	Precedence		
Status	Routine	Priority	Urgent
In-patient	6291	169	142
Out-patient	616	1	0

directly from Scott. Priority and urgent destinations are determined by the doctors involved. Out-patients traveling to follow-up examinations and treatment not requiring hospital admission are not regulated or validated by ASMRO, even for routine precedence. In total, 928 movements (fourteen per cent) were not regulated, and 617 out-patient movements were neither regulated nor validated.

In terms of planning, urgent patients require the greatest amount of planner involvement. 142 urgent patients required the launching of 56 special missions dedicated to, in some instances, a single patient, or the rerouting of aircraft assigned to routine missions. To the maximum extent possible, planners attempt to reroute aircraft to pick up urgent cases, balancing the needs of the patients already onboard an aircraft flying a scheduled route and those scheduled but yet to be picked up against the needs of the urgent patient.

The data also allowed us to classify patients by origin. Of particular interest are those who come from overseas origins to destinations in the US. 938 patients (thirteen per cent) arrived at either Travis AFB or Andrews AFB from Alaska, Hawaii, US territories and foreign countries. Flights to the US are scheduled and very stable, so in a sense, they generate scheduled movements within the US. This means that planners can assume pickup requirements at Travis and Andrews Air Force Bases will always exist on specific days of the week, which aids the scheduling process. Overseas patients do present one unique planning problem. Of 842 patients who moved between non-adjacent regions, 506, or sixty per cent, were from overseas origins. Roughly half of the 506 traveled from Andrews AFB to Texas or California, and another 101 from Travis AFB to the Northeast.

Virtually all psychiatric patients and all drug abuse patients were active duty personnel. As we mentioned earlier, these patients are usually treated at facilities operated by their parent service. 1512, or twenty-one per cent of all patients were in these two major classes.

The litter, ambulatory and troop classes imply different levels of enroute care. Patients in troop status do not require extensive care enroute, if any, and would be able to evacuate an aircraft in an emergency without assistance. Litter patients imply significantly greater care needs, and larger medical crews must be assigned to a flight to ensure that litter patients would be evacuated in emergencies. Enplaning and deplaning patients in the various classes will differ both in the amount of time required, and in the number of crew members or ground crewmen required.

Because they are patients, moving tnem often involves more than is apparent in the data. Their individual circumstances can greatly complicate transportation planning. One patient's illness may require a quarantine, barring the carrying of any other patients. Special medical needs, for example, a reduced cabin altitiude for a thoracic surgery patient, may impose a direct routing restriction from origin to destination for that patient, reduce aircraft range and hence the choice of destinations because of the greatly increased fuel consumption at low altitude, or add additional stops for refueling.

One important observation we made when collecting the data was an historic lack of automated patient information processing support. Until recently, the only computer available to the organization compiled statistical summaries for upward reports using historical data extracted from handwritten movement manifests. In fact, while patient movement planners now have access to an automated data base on patients currently being served, that system was designed only to record and retrieve data, and not to function directly as a decision support system.

2.3.2 Patient Flow Model. Human interaction in space requires both people and places. Having briefly described the demographic charactersitics of the system's patients, we now examine the places, and the interactions among them. We will attempt to establish as comprehensively as we can the structuring of flows between pairs of and among the whole system of DOD medical treatment facilities.

We have several purposes in mind. We want to identify factors influencing origin-to-destination trip production so that we can better represent the patient demand schedule in our system model. By a process of abstraction, we will use a number of analytical techniques to isolate patterns of movement and construct a flow model. Potts and Oliver recommend that we develop such models to

1) Represent manifestations of individual and mass propensities for interaction;

2) Demonstrate the nature and complimentarity of supply and demand;

3) Indicate those nodes whose characteristics --size, dominance, centrality, etc. -- should be better understood;

4) Extract the underlying flow structure, its recognizable order and organization, channeling, and regionalization;

5) Allow us to relate flow to the transportation network and external factors (capacity, cost, etc.). [POTT72]

Because we were not given individual patient diagnoses, the first two capabilities cannot be achieved completely. We have focused our attention on the last three. The last capability will be demonstrated in Chapter 6.

We will use four means to construct our patient flow model. First, we will employee aggregate descriptive statistics to characterize the major flow propensities.

Then, we will analyze nodal hierarchy, by examining the functional importance of individual nodes. Using dyadic factor analysis we will then attempt to explain the network flow structure. Our theoretical rationale for this is that to the extent that the nodal structure is hierarchical, we would expect the flows to be structured hierarchically also. Finally, we will look at the generation of movement requirements over time.

In Table 2.10, we see that of the 500 hospitals from which patients trips originated or were destined, hospitals in San Antonio, TX, alone accounted for nearly one-fifth of all patient destinations. San Antonio and Washingtion, D.C. facilities admitted one-third of all patients moved, and with Fitzsimmons Army Hospital in Denver and the US Air Force Regional Hospital at Keesler AFB in Biloxi, MS, approximately one-half of all patients transported in the domestic system. Nine groups account for over two-thirds of all destinations, and for the fifteen hospital groups listed, the percentage increases to over eighty per cent. The colocation of Aeromedical Staging Facilities at bases serving six of the first seven most frequently used destination hospital groups is therefore not surprising.

A similar dominance in originating hospitals (Table 2.11) is also present but less pronounced, primarily because the aeromedical system emphasizes moving patients from many smaller medical facilities to a few, very large

TABLE 2.10

RANKINGS OF DESTINATION HOSPITALS

Hospital Group[a]	Servicing Airfield	Number of Patients	Percentage of All Patients
1. Wilford Hall MC Brooke Army MC	Kelly AFB, San Antonio, TX[b]	1406	19.21
2. Walter Reed Army MC National Naval MC Bethesda Malcom Grow MC	Andrews AFB, Camp Springs, MD[b]	1096	14.97
3. Fitzsimmons Army MC	Buckley ANGB, Denver, CO[b]	756	10.33
4. USAF MC Keesler	Keesler AFB, Biloxi, MS[b]	435	5.94
5. David Grant MC	Travis AFB, Fairfield, CA[b]	407	5.56
6. Wright-Patterson Regional MC	Wright-Patterson AFB, Dayton, OH	320	4.37
7. Scott AFB Regional MC	Scott AFB, Belleville, IL[b]	307	4.19
8. Portsmouth Naval Regional MC	Norfolk NAS, Norfolk, VA	233	3.02
9. San Diego Naval Regional MC	Mirimar NAS, San Diego, CA	221	3.18
10. Eisenhower Army MC	Hunter AAF, Augusta, GA	180	2.46
11. William Beaumont Army MC	Biggs AAF, El Paso, TX	179	2.45
12. USAF Hospital McChord	McChord AFB, Tacoma, WA	169	2.31
13. Naval Regional MC Jacksonville	NAS Jacksonville, Jacksonville, FL	145	1.98
14. USAF Regional Hospital	Sheppard AFB, Witchita Falls TX	117	1.60
15. USAF Regional Hospital Eglin	Eglin AFB, Valpariaso, FL	94	1.28
	Totals	6065	82.84

[a] Hospitals served by the same airfield.

[b] An Aeromedical Staging Facility is located here.

medical centers. Hospitals in Florida, region 5, and the Southwest are the largest patient generators.

The relative importance of these facilities within their respective regions varied. In region 3, 91.27 per cent of all patients moved went to Fitzsimmons Army Hospital, while only 33.47 per cent of all region 2 patients went to the USAF Regional Hospital at Keesler AFB. The San Antonio hospitals were also the major regional facility for region 5 patients, serving 74.35 per cent of all paitents sent to region 5 hospitals.

Table 2.12 shows that, while fewer than half of all patients were moved intraregionally, the combination of intraregional movements and interregional transfers to fourteen major facilities accounted for 90 per cent of all trips. Of all 7321 patients moved, only 842 (11.5 per cent) were moved to non-adjacent regions, which constituted only 20.75 per cent of all interregional transfers. As we observed earlier, the majority (506) of those non-adjacent transfers were patients with origins outside the US. One-half of that group were sent to the major facilities in Texas and Washington, D.C.

Table 2.13 examines interregional transfers from a slightly different perspective. Given the six regions, there are 30 possible transfers between regions. The first six listed account for over half of all interregional

TABLE 2.11

RANKINGS OF ORIGINATING HOSPITALS

Hospital Group[a]	Servicing Airfield	Number of Patients	Percentage of All Patients
1. Walter Reed Army MC National Naval MC Bethesda Malcom Grow MC	Andrews AFB, Camp Springs, MD	876	11.97
2. David Grant USAF MC	Travis AFB, Fairfield, CA	502	6.86
3. Wilford Hall MC Brooke Army MC	Kelly AFB, San Antonio, TX	266	3.63
4. USAF Regional Hospital Eglin	Eglin AFB, Valparaiso, FL	207	2.83
5. USAF Hospital Patrick	Patrick AFB, Cocoa Beach, FL	194	2.65
6. USAF Hospital Homestead	Homestead AFB, FL	171	2.34
7. USAF Hospital Pope	Pope AFB, Fayetteville, NC	164	2.24
8. Scott AFB Regional MC	Scott AFB, Belleville, IL	161	2.20
9. Fitzsimmons Army MC	Buckley ANGB, Denver, CO	157	2.14
10. USAF Hospital Luke	Luke AFB, Glendale, AZ	154	2.10
11. USAF Regional Hospital MacDill	MacDill AFB, Tampa, FL	145	1.98
12. USAF Hospital McChord	Mountain Home AFB Mountain Home, ID	144	1.97
13. Irwin Army Hosp Ft Riley	Salina Municipal Salina, KS	136	1.86
14. Ft Campbell Army Hospital	Ft. Campbell AAF Hopkinsville, KY	127	1,73
15. USAF Hospital Ellsworth	Ellsworth AFB, Rapid City, SD	108	1.48
	Totals	3512	47.97

[a] Hospitals served by the same airfield.

TABLE 2.12

PATIENT TRANSFERS

Patient Group	Number of Patients Transferred	Percentage of All Patients Transferred
Intra-regional transfers (Origin and destination hospitals in the same region)	3162	43.80
Inter-regional transfers to destination hospitals served by airfields with an ASF:		
Kelly AFB	882	12.22
Andrews AFB	797	11.04
Buckley ANGB	223	3.09
Scott AFB	191	2.65
Travis AFB	142	1.97
Keesler AFB	36	0.50
Inter-regional transfers to destination hospitals served by airfields without an ASF:		
Wright-Patterson AFB	180	2.49
Gray AAF	146	2.02
Norfolk NAS	143	1.98
McChord AFB	113	1.57
Mirimar NAS	107	1.48
Jacksonville NAS	100	1.39
Lawson AAF	96	1.33
Sheppard AFB	67	0.93
Totals	6498	90.01

transfers, ten for two-thirds, and each of these combinations involved adjacent regions. Of the 2059 patients involved in the first six transfers, 671 were destined for San Antonio hospitals. 149 of those were overseas patients moved from Andrews AFB. 343 of the 363 patients moved from region 2 to region 5 went to San Antonio. Of the 350 region 4 to region 5 transfers, 317 went to either San Antonio, Wichita Falls or El Paso, 170 from region 4 to San Antonio alone.

TABLE 2.13

INTER-REGIONAL PATIENT TRANSFERS[a]

From Region	To Region	Number of Patients Transferred	Percentage of All Transfers	Number with Destination at an ASF
2	1	672	16.56	551
2	5	369	9.10	343
4	5	350	8.63	170
1	2	239	5.89	13
1	6	232	5.72	25
1	5	197	4.86	158
6	1	196	4.83	155
6	5	157	3.87	132
3	4	153	3.77	44
1	4	142	3.50	46
4	1	130	3.20	65
6	3	117	2.88	100
5	2	116	2.86	16
3	5	113	2.79	79
4	2	103	2.54	5
	Totals	3286	81.00	1902

[a] Recovered patient and medical attendant transfers are excluded.

This preliminary examination of relative nodal functional importance and flow propensities leads us to hypothesize that patient movement follows a hierarchical pattern of intraregional flows between small facilities and into regional centers, and between interregional centers, among nodes that are also hierarchically ordered. To test these propositions, we first examined the data for evidence of hierarchical structure or ordering among individual nodes Of 215 civil airports and military bases that served as either an origin, destination, or both, 54 appeared at least 40 times in patient records, and in all but 60 of 7321 records. In other words, only 60 origin-destination pairings did not contain at least one member of the set. 6345 patient origins and 6839 destinations, 86.7 and 93.4 per cent of all patients respectively, were bases in this set. 7261, or 99.2 per cent, of all patient movements began or ended at one of the 54 bases, and in 5923 (80.9 per cent) cases, both the origin and the destination were part of the set.

Comparing the largest and smallest entries in Tables 2.10 and 2.11, it is evident that there is considerable difference in the size of nodes as measured by the numbers of patients they send to or receive from other facilities. Except for Kelly, Scott and Andrews Air Force Bases, the rankings and membership of the largest origin and destination hospital groups are different, and the relative magnitudes are large between the first and fifteenth ranked hospitals. With several hundred different hospitals, there must be many hospitals sending or receiving very few or even single patients. Figure 2.10 illustrates this preponderance of low-volume nodes and the relative scarcity of high-volume nodes.

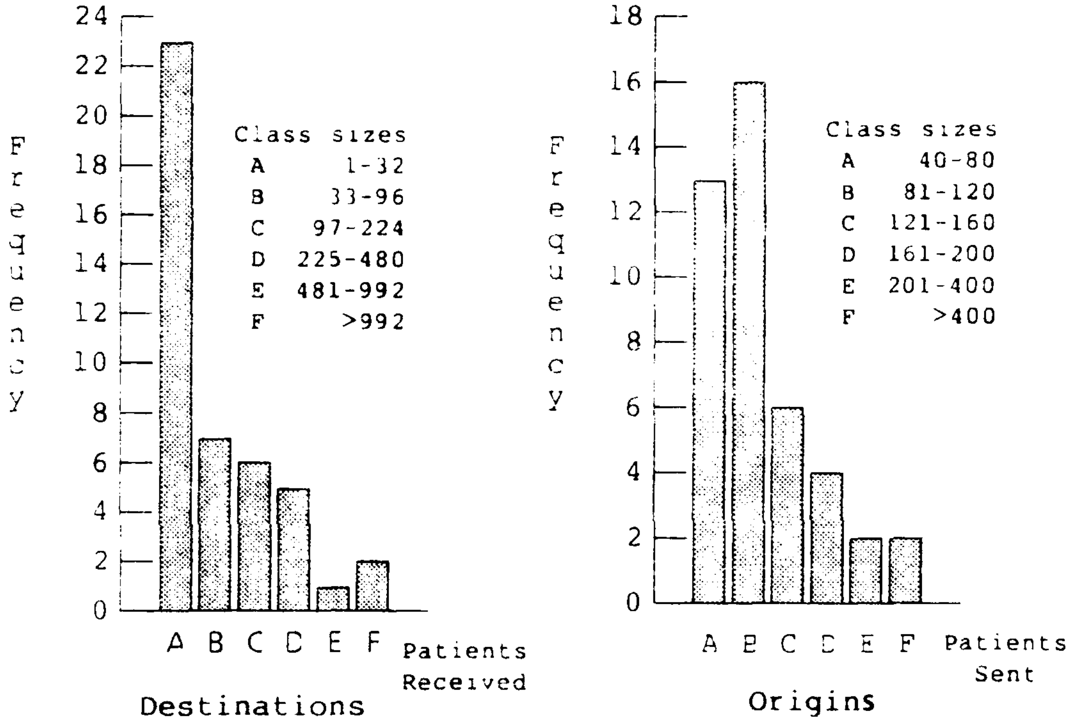


Figure 2.10. Patient origin and destination frequencies.

Lowe and Moryadas assert that the principal determinants of the volume of interaction (flow) between nodes are their size (as either origins or destinations) and the distance separating them. [LOWE75] We are principally concerned with the first. More specifically, we need to establish the flow propensities between nodes, and the flow patterns. We will use the methods that find the <u>dominant</u>

and salient flows to establish flow propensity, and factor analysis to extract the structural patterns.

The method of dominant flows is widely used to establish a dominance order among nodes. If we define f_{ij} as the actual flow of patients from facility i to facility j, $F = [f_{ij}]$ as the matrix of all flows, then define

$$F_j = \sum_i f_{ij} \tag{2.1}$$

the total flow into facility j, to be the functional importance of node j. For any i, if the largest flow in row i is f_i^*, and $F_i < F_j$, then f_i^* is a dominant flow.

When the dominance technique is applied to flows among the 54 major hospital groups, the graph in Figure 2.11 indicates the nature of the dominance order and channeling of flows. Separate subsystems based on flows into the region 3 and 4 centers (Buckley and Travis) result in a partitioning of flows in the system. The flow from GFA to TCM in the Pacific Northwest is subordinate to the flow from TCM to Travis. The largest subsystem of dominant flows links four regional centers, Andrews, Kelly, Scott and Keesler. Within that subsystem, two centers, Scott and Keesler, are subordinate to Kelly and Andrews, given the high flows (1406 and 1096) into the latter two nodes.

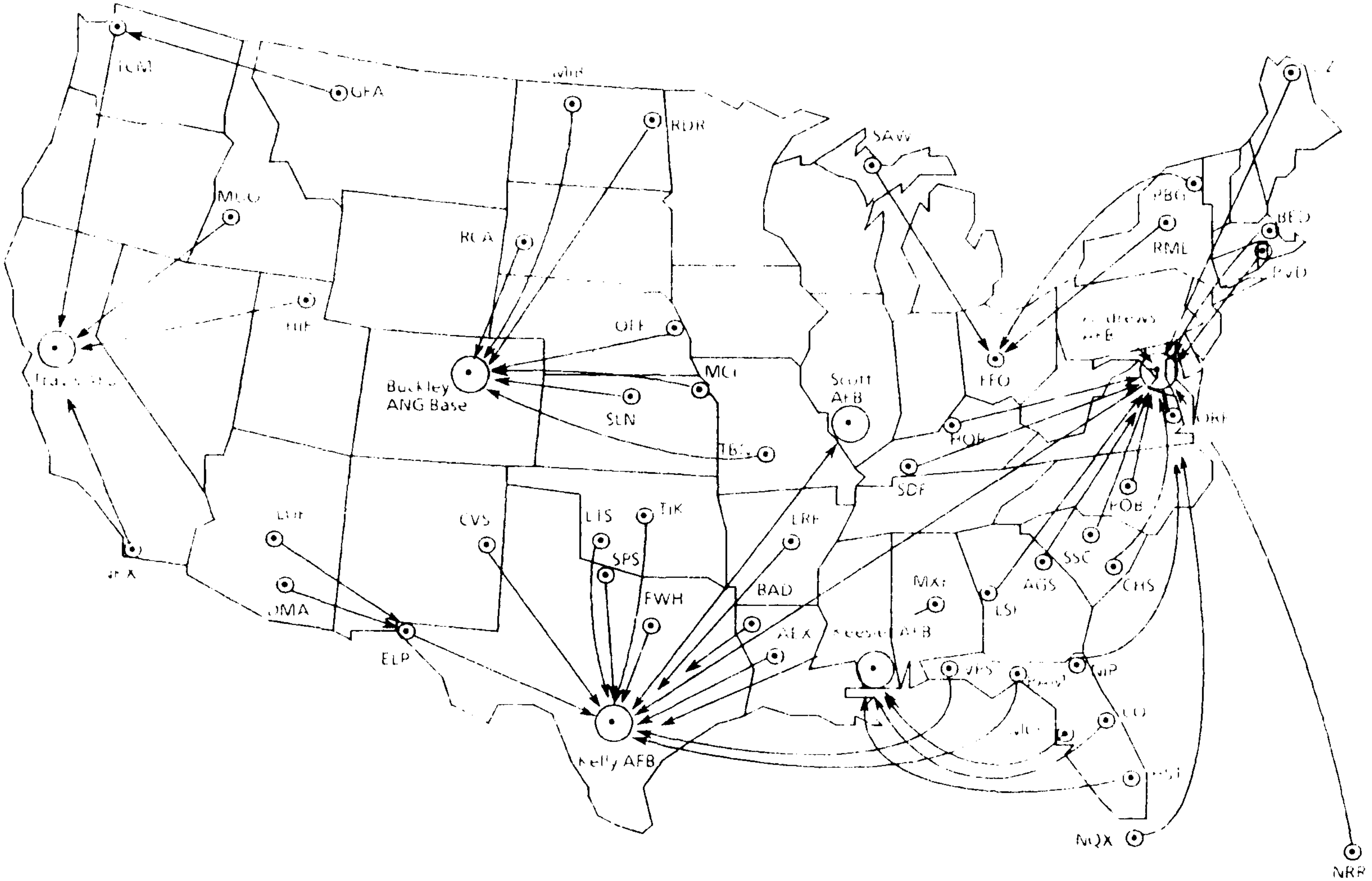


Figure 2.11. Dominant patient flows.

The Southeastern US is effectively subdivided into subordinate flow networks feeding the two largest nodes. Also note the further subordination of flows into ELP and BIX. Underlying three patterns (TBN-BKF, POB-ADW, and LUF-ELP) is a standing arrangement recognized by ASMRO between two hospital commanders for the transfer of patients between their two facilities. In each instance, these patterns are aligned with interregional trunk routes.

Soja [LOWE75] developed a second method of flow propensity that identifies salient flows, those that are significantly higher than expected, based on the r-by-c contingency table statistical technique. He defines the the salience measure of flow f_{ij} as

$$ra_{ij} = \frac{f_{ij} - e_{ij}}{e_{ij}} \qquad (2.2)$$

where

$$e_{ij} = \frac{\sum_{i} f_{ij}}{\sum_{i} \sum_{j} f_{ij}} \sum_{j} f_{ij} \qquad (2.3)$$

in which the fractional term is the proportion of all patient flows destined for node j, and the second term is the total flow originating at node i. That is, if node j receives x per cent of all patients, then that same proportion should originate at all origins. The resulting flow hierarchy identified by the salience technique is shown in Figure 2.12.

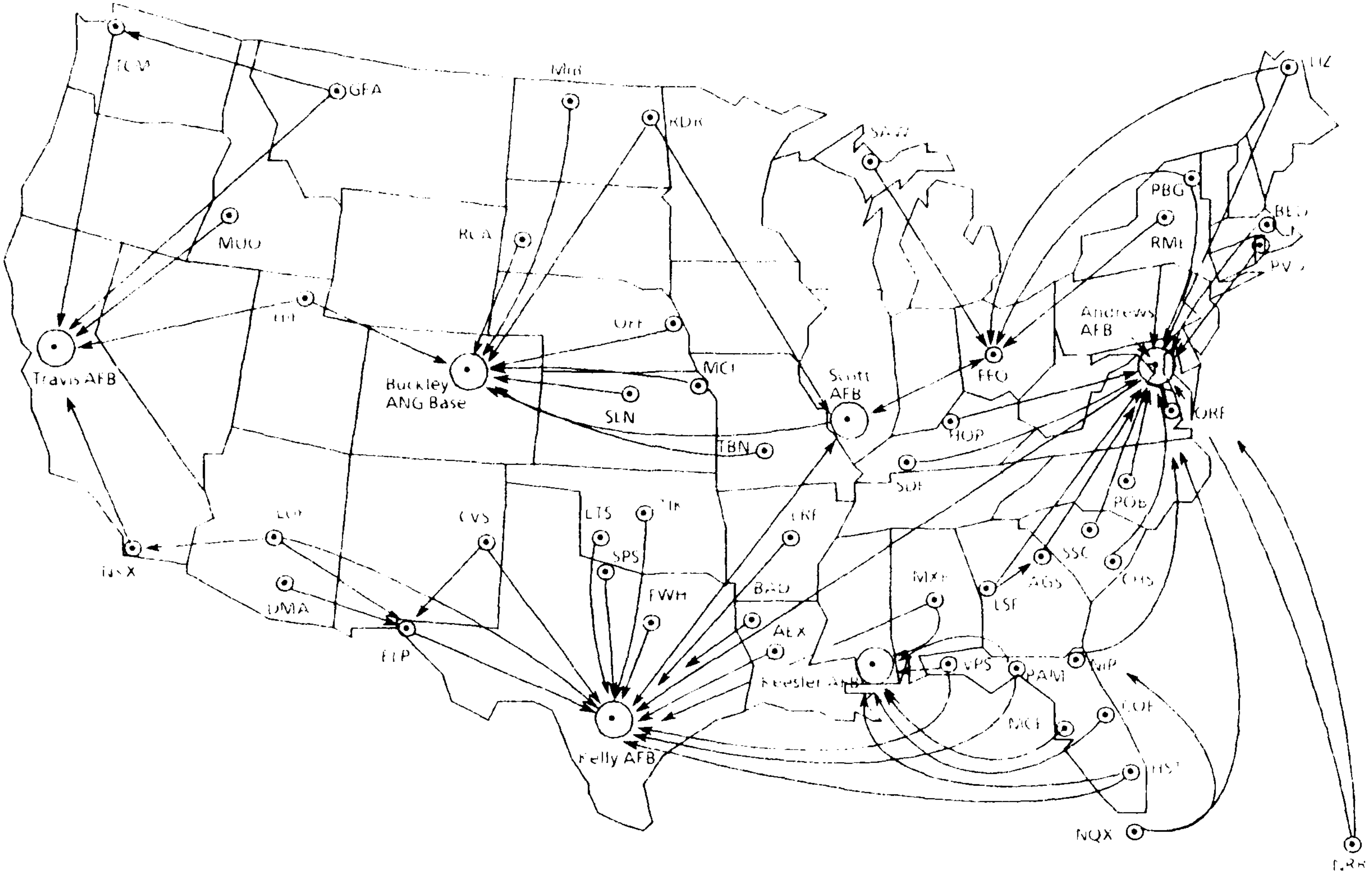


Figure 2.12. Salient patient flows.

Dominance and salience analyses strongly suggest that patient flow propensities are regional. To find the regional patterns, we use dyadic factor analysis. [LOWE75] Dyads are groups of origins and destinations. Since the flow matrix contains all flow transactions between all origin (row)-destination (column) pairs, by applying factor analysis on the flow matrix, functional regions can be identified based upon similarities in flow patterns.

Factor analysis uncovers the r common factors (or components) that "explain" the flow data and account for the maximum possible variance. Where salience and dominance are descriptive, factor analysis is explanatory; the factors are assumed to be causal in nature, scientifically replicable and of theoretical interest, determining the correlation among variables. [IMSL82] Each factor is a group of nodes among which there is maximum homogeneity. Between factors groups, there is maximum heterogeneity; for maximum explanation, the factor-analytic routine uses orthogonal rotation to reduce correlation among factors to zero. Total initial variance is iteratively reduced in diminishing amounts as residual correlation is computed and each factor is extracted. The variances accounted for by each factor, then, are additive.

Initially, the factors are devoid of any substantive meaning. The principal analytical task is to interpret the results in terms of real phenomena associated with the original variables. [BASI83] To do this, high factor loadings and factor scores are taken to mean that a high correlation exists between the orginal variables and those scores. We intrepreted the factors to be characteristic linkages between hospitals based on similarities (common origins and destinations) and intensities (volumes) of flows. Based on our dominance and salience maps, we hypothesized that the linkages would delineate functional regions within the US.

Table 2.14 reports the total variance explained by the factor analysis using the IMSL Library on a CRAY-1S computer. Of the total variation of 54.0023, 48.3083 (89.46

TABLE 2.14

AEROMEDICAL PATIENT FLOW DYADIC FACTOR ANALYSIS RESULTS

		Per Cent
Total Variance	54.0023	100.00
Variance Explained by Each Basic Pattern		
1	13.9866	25.90
2	9.6297	17.83
3	6.4314	11.91
4	5.5584	10.29
5	3.5042	6.49
6	2.1413	3.97
7	1.7300	3.20
8	1.4591	2.70
9	1.3187	2.44
10	1.2525	2.32
11	1.2964	2.40
Totals	48.3083	89.46

per cent) is explained by the eleven most signif-icant factors. The first five components appear to be regionalized flows into the regional centers, which had factor loadings 5 or more standard deviations from the mean loading. Factor scores were greater than 0.5 for all origins connected to them in Figure 2.13. Some exceeded 0.9. Relative loadings for the first five patterns reflect a nearly inverse linear relationship with the number of origins highly correlated with each major destination. The sixth factor is flow from two Arizona hospital groups into El Paso, TX, one of the so-called approved local agreements recognized by ASMRO as exempt from the nearest facility rule. In factors seven through eleven, origin-destination relationships are reversed. (We could not interpret the eighth factor because no factor scores exceeded 0.3).

Based on these results, we conclude that the dominant patient flows are highly regionalized and interconnect the most dominant nodes within the nodal hierarchy. Strong internal linkages appear to exist between origins and destinations. Further, these findings correspond directly with our initial descriptive observations of the relative importance of a few major centers as destinations, with relatively more significant origins. The partitionings indicate that patient movements are not concentrated in a single geographical area, but are dispersed through six or seven major subregions extending over almost all of the US.

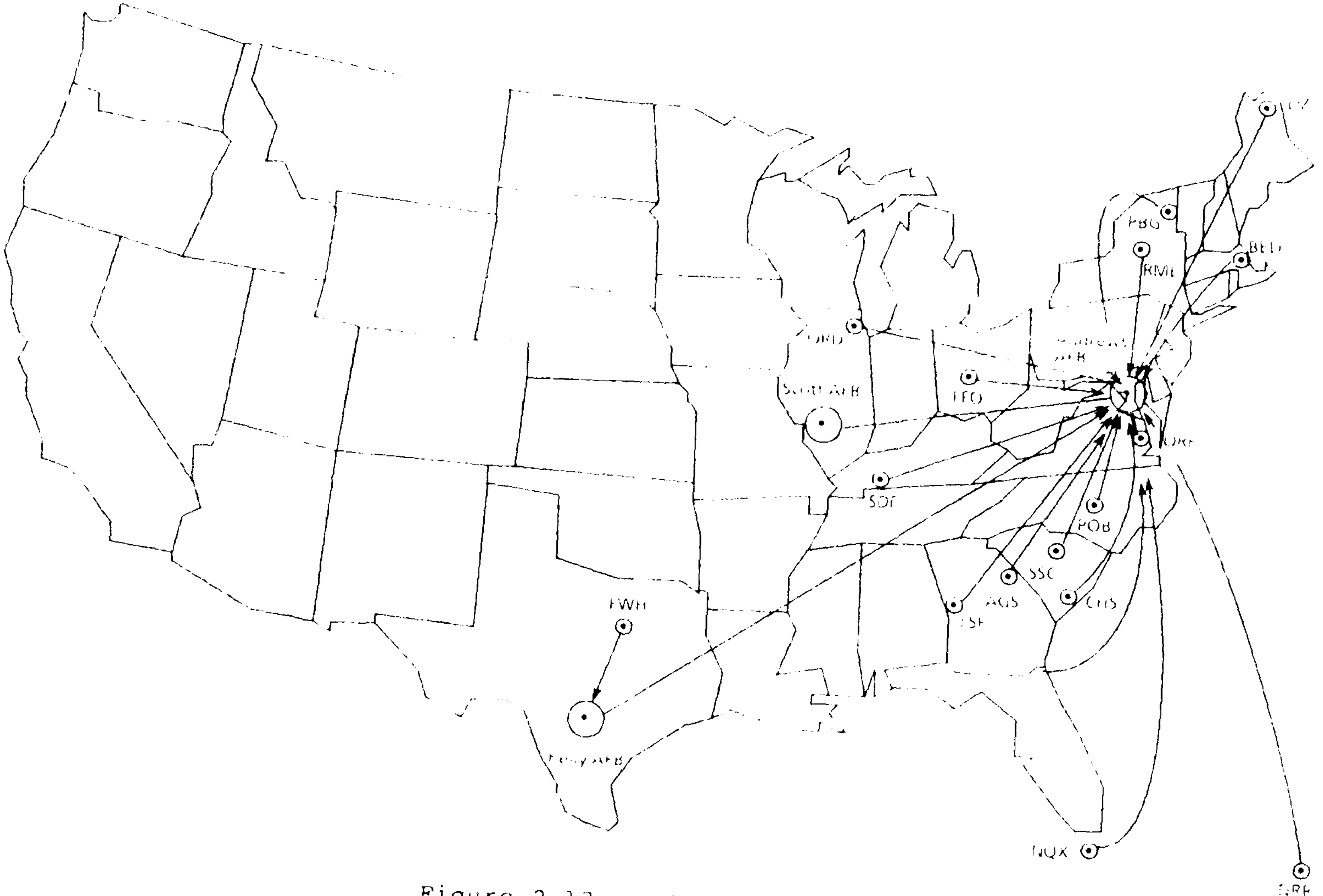


Figure 2.13a. Factor #1.

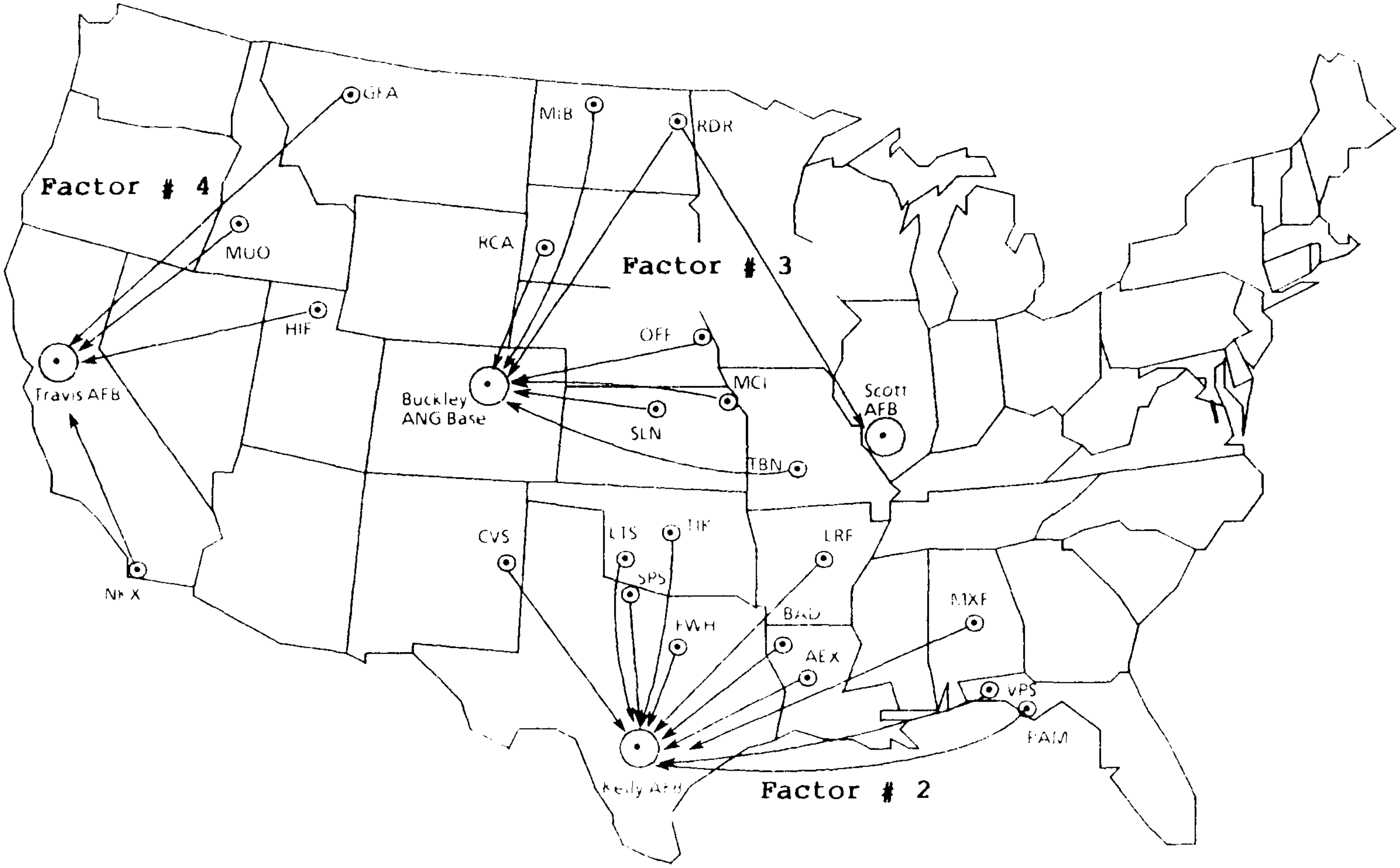


Figure 2.13b. Factors #2, 3# and 4#.

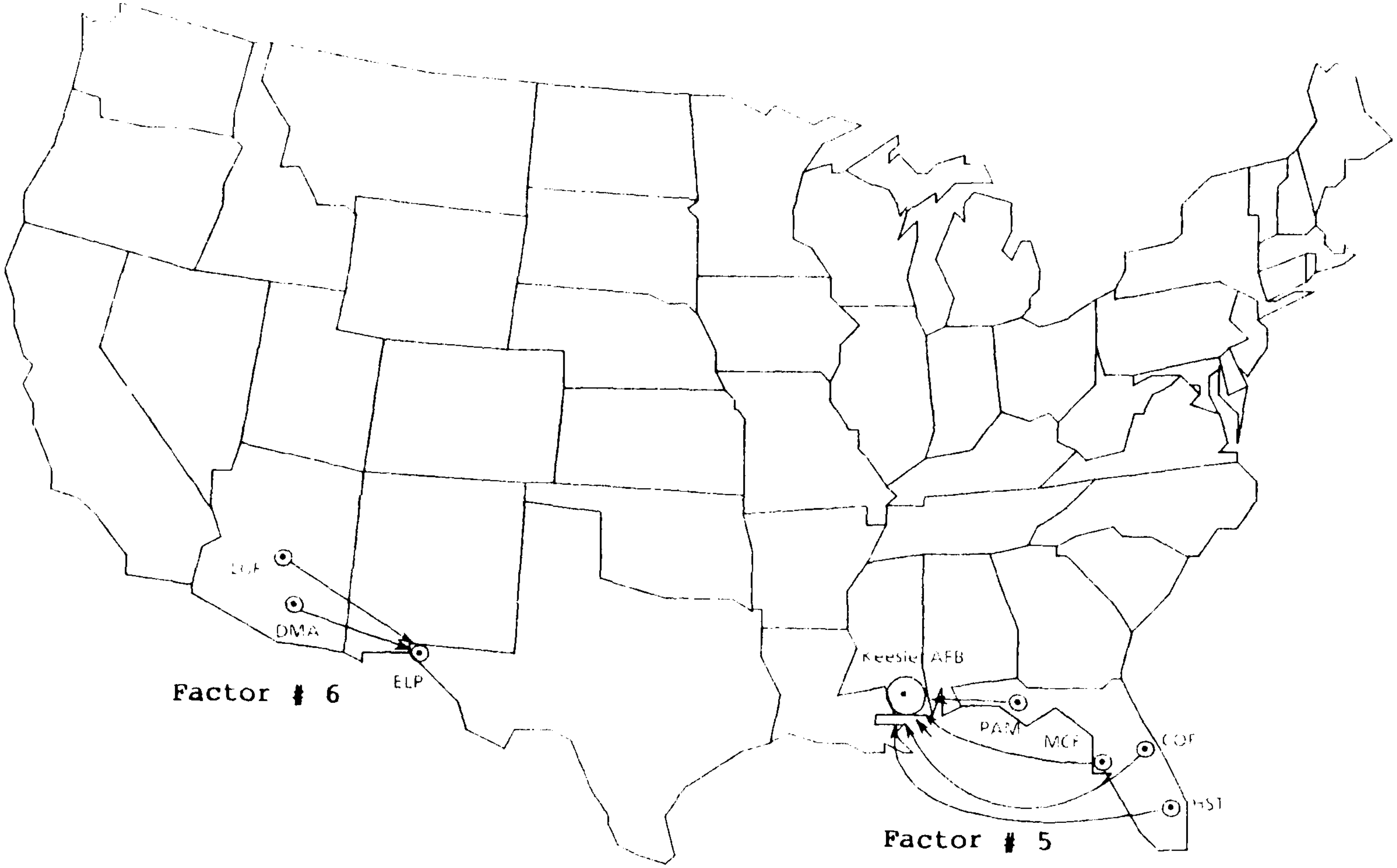


Figure 2.13c. Factors #5 and #6.

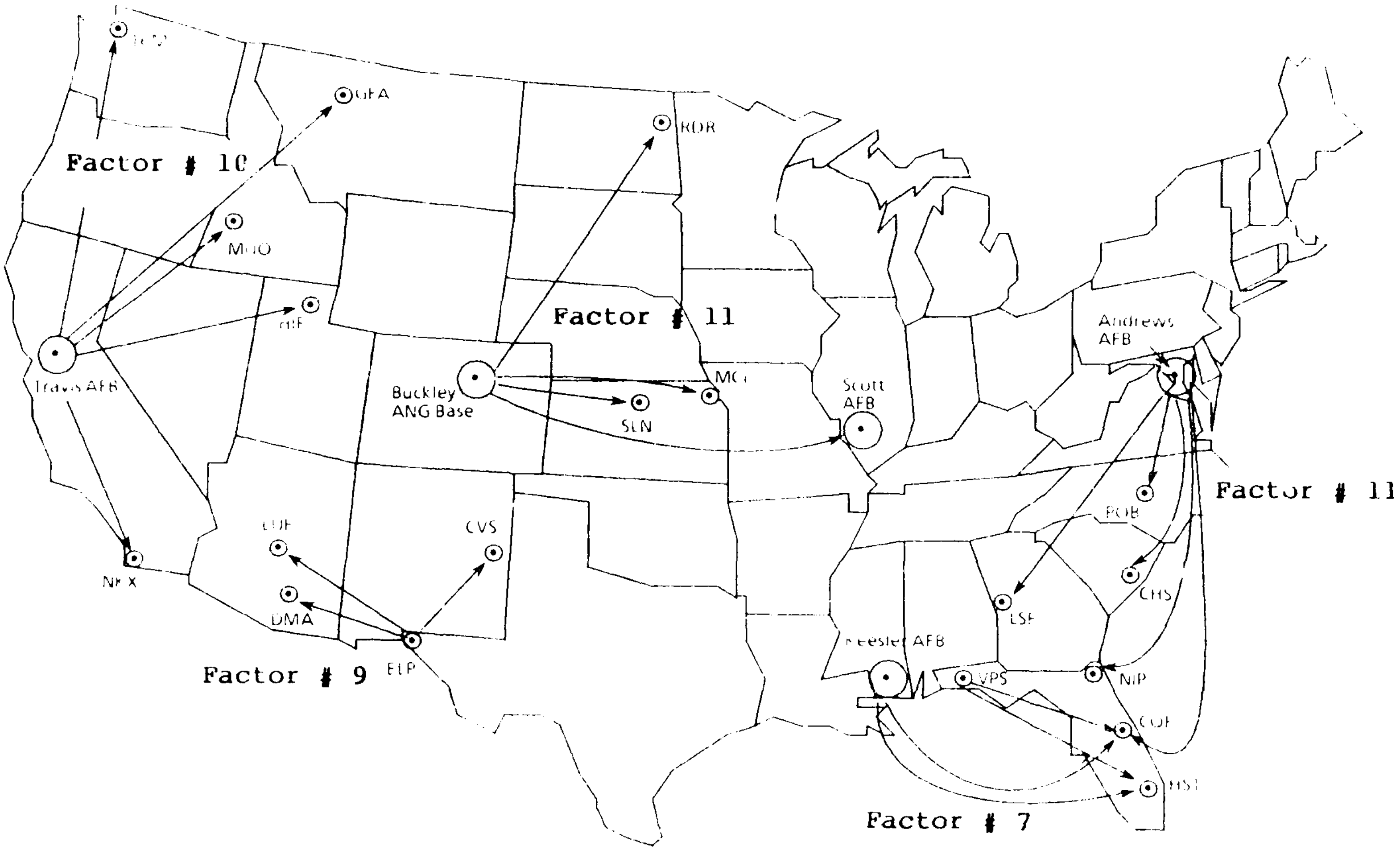


Figure 2.13d. Factors #7, #9, #10, and #11.

To establish the chronological patterns of patient flow, we categorized patient movement by the first day the patient could be moved. We were particularly interested in the variety of daily movement demands, in terms of the number of discrete stops involved (since this number is limited by fleet size, operating rules governing the length of the crew duty day and the maximum number of stops per route allowed, route segment lengths and ground stop durations), and the pattern of demand over the days of the week. We hypothesized that midweek and late week demands would exceed the maximum number of stops that could be made that would provide one-day service to every patient.

TABLE 2.15

DAILY PATIENT MOVEMENT DEMAND

	Day of the Week				(Demand/Stops)		
Week	Monday	Tuesday	Wednesday	Thursday	Friday	Saturday	Sunday
1			84/62	106/60	73/57	16/14	6/9
2	70/48	90/60	82/64	112/72	23/29	15/17	14/15
3	60/45	97/63	81/56	113/58	79/59	19/20	20/18
4	76/52	85/61	60/52	25/24	63/49	18/20	33/21
5	79/52	114/60	75/52	100/53	76/56	20/21	12/16
6	95/58	83/57	89/64	87/57	71/55	17/19	26/22
7	65/44	84/66	70/54	100/64	65/56	20/24	21/22
8	72/52	78/58	53/59	67/42	29/29	16/17	9/12
9	10/14	34/35	44/41	44/39	24/26	21/22	12/13
10	13/16	61/51	83/54	91/56	58/42	21/20	32/30
11	87/52	80/56	78/44	116/62	76/56	27/27	31/28
12	75/22	88/55	63/49	104/55	91/61	15/16	30/25
13	86/53	65/48	71/51	118/63	76/58	20/22	38/36
14	83/45	62/34					
Demand	77.78	84.78	77.00	106.2	70.00	18.78	22.00
Stops	43.33	57.33	54.22	60.00	52.44	20.67	23.56

In Table 2.15, the entries show for each day the number of distinct origin-destination movement combinations (left) with the number of different bases in those combinations. That is, only one movement was counted even if several patients required transfer between the same two bases on the same day. And, if the same base served as both an origin and a destination, it was only counted once. This provides an indication of how many stops were required on that day to make all transfers.

To test the null hypothesis that mean daily demands are tne same, a one way analysis of variance procedure was used. Because Thursday of week 5 and the Mondays in weeks 10 and 11 were holidays (Thanksgiving, Christmas and New Years respectively), demands in those weeks were clearly much lower. These and the partial weeks (1 and 14) were omitted. The tabulated F value to test the null hypothesis of no difference in daily means at a level of significance of $\alpha = .05$, 6 degrees of freedom among the daily samples, and 56 degrees among replications, was $F_{.05}$ $(6,56) = 2.28$, and $F_{.01}$ $(6,56) = 3.16$. [MEND68] It is clear that the computed F value of 66.3 exceeds both tabulated values and thus falls in the rejected region. We therefore reject the null hypothesis and conclude that a significant difference exists between the daily mean number of patients requiring movement. Using Tukey's ω-test to compare all pairs of daily means to test the null hypothesis of mean equality, the tabulated q statistic for 7 days and 56 degrees of freedom is 17.33, with a probability of falsely rejecting the null hypothesis on at least one comparison of α = .05. Grouping the means in the following way to show pairs that differ by more than 17.33,

18.78	22.00	77.78	84.78	77.00	70.00	106.2
[Sat	Sun]	[Mon	Tues	Wed	Fri]	[Thurs]

we would reject the null hypothesis for pairs that include Saturday or Sunday and any weekday, Thursday and any other day, and the other four weekdays and any other days. We can conclude, then, that weekday demands are significantly higher than on weekends, and that Thursdays produce particularly high movement requirements.

To conclude, our evaluations of patient flows suggest several salient characteristics of patient movement. Flow hierarchy is related to the nodal hierarchy; the largest nodes in terms of numbers of patient origins and destinations are involved in the largest flows. The structural flow patterns reflect a strong regionalization coupled with large flows between region centers, and they cover virtually all of the continental US. Fewer destinations have relatively importance than many more relatively less important destinations. And finally, there are pronounced differences in demand between days of the week, with the number of stops required greater than the maximum possible under current mission levels and rules.

2.3.3 Analysis of the Air Route Network. The last section addressed the structure of patient movement demand, but not how that demand was actually met. Patient transportation needs focus on origin and destination medical facilities locations. Movement needs are met by assigning aircraft to routes interconnecting patient origins and destinations. Patient flow and aircraft routing structures are certainly related, but not identical, since the routing structure must also include points where patients can be transferred from one aircraft to another and where patients, aircraft and crews can be accomodated overnight. The purpose of this section is to examine the aircraft routing structure used to satisfy the demand just described.

The function of a route is to contribute to movement efficiency by structuring flows. By directing and agglomerating individual patient movements, the routing structure greatly reduces the distances that would have to be traveled to serve each individual separately. Different route configurations can yield widely varied efficiencies, and the presence or absence of alternative routes can greatly influence how directly a patient reaches his or her destination. Because patient transfer demand occurs over time, the frequency of travel over routes determines how often each node is served, and, therefore influences how quickly the patient is served. From the last section, we know that some nodes generate high demands and receive disproportionately high levels of patients for treatment. The essential impact of a routing structure, then, is to enhance the relative accessibility of the functionally more important nodes on the network. We would expect a routing system to consist of high capacity, high frequency routes between larger nodes (and connecting additional nodes along the path between them), direct links to smaller nodes in close proximity to the larger nodes, and subnetworks interconnecting smaller nodes in close proximity to each other. Finally, the number of routes should vary inversely with the volume of flow and distance or length, and directly with node density in a geographic area.

In this section we will analyze the aeromedical aircraft routing structure from aircraft data gathered during the same period in which the patient movements described in the last section occurred. Graphically, Figure 2.14 shows the route segments that were flown most frequently. Knowing that the central base is Scott AFB, that crews and aircraft must pass through the central base frequently, and from the last section, that destinations are most frequently the regional centers and a few major medical facilities, the trunk system and the resulting circulation patterns are apparent. While useful, this depiction contains only about one-third of the total number of route segments, and omits a considerable portion of the structure. A simple frequency count revealed that over 855

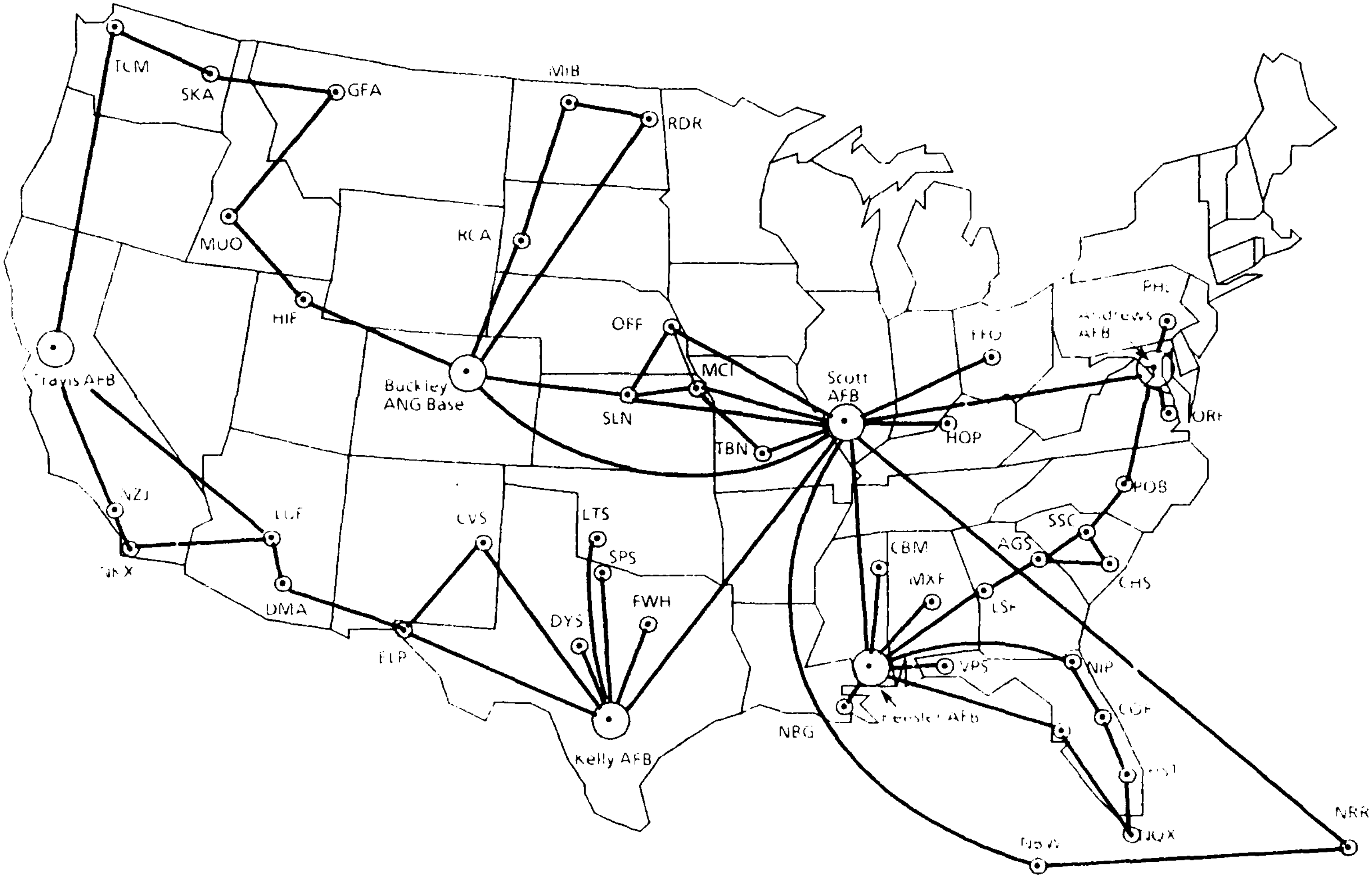


Figure 2.14. Most Frequently Traveled Route Segments.

segments were flown once, and another 213 only twice. The resulting graphic depiction with these segments added would be uninterpretable, yet still omit the remaining third of all segments flown. Beyond providing an overview of the structure, the route structure diagram does not adequately provide insights into the underlying network structure.

The fact remains, though, that one of the principal methods in route structure analysis is constructing and analyzing maps. Maps are very useful in presenting spatial relationships, but they vary considerably in the amount of information they contain. Usually, but not necessarily, they are two-dimensional projections, with one or at most only a few metrics. Mapping involves considerable choice, selecting the type of projection; depicting route locations (and their absence); differentiating route segments by type, function, purpose, or quality; measuring length, capacity, or other salient characteristics; displaying and interpretating relationships, such as intersections; and showing the relative importance or value of network entities. In short, choosing what and how to depict involves more judgement than science, a consistent theme of the systems approach.

The closest that we have found to a standard for analyzing and interpreting transportation network route structures is linear graph theory, a branch of combinatorial topology. Through considerable abstraction, a graph-theoretic approach maps a route structure into a minimal set of characteristics. Boundaries, scale, proportion, and most other information are lost, and the resulting representation often introduces meaningless new structures, such as the mid-air intersection of route segments. And, while preserving spatial sequences, temporal relationships are lost. Because the current mode of operation is demand-responsive routing, changes over time are the norm, and not the exception. On the other hand, when faced with making sense out of thousands of segments connecting several hundred places, we can usefully exploit graph-theoretic techniques to reduce the data to something meaningful as one means to better understanding.

In our discussion of route structure and routing problems, we will make use of the following concepts. Given the set of nodes $N = \{1,2,\ldots,n\}$, $n = |N|$, we define the directed arc from node i_1 to node i_2, representing the direct connection of two places, as the ordered pair, (i_1,i_2), where i_1 and i_2 index the initial and final (or terminal) nodes of the arc. Given the set of all arcs, $A = \{1,2,\ldots,a\}$, $a = |A|$, the graph $G = \{N,A\}$. We assume there exists a function d:A-->C, where the matrix $C = [c_{ij}]$ and c_{ij} is the cost of traveling from city i to city j.

A path is an ordered sequence of directed arcs

$$P = ((i_0,i_1),(i_1,i_2),\ldots,(i_{k-1},i_k)), \qquad (2.4)$$

where the final node of one arc is the initial node of the next arc. An elementary path has no repeated nodes, and in an elementary circuit, $i_0 = i_k$. If a path exists between them, i_k is accessible from i_j. If a graph is connected, each node is accessible from every other node. In a complete graph, directed arcs connect every node pair. A subgraph contains nodes that are not connected with other nodes; a graph may consist of several disconnected subgraphs. A spanning subgraph contains all nodes, but not all arcs of G. In planar graphs, arcs only intersect at the nodes. Since graphs of air route networks are virtually certain to intersect at points other than nodes, we assume they are non-planar.

The number of arcs entering or leaving (incident to) a node, the ratio of arcs to nodes, and other characteristics of a graph can be expressed by a set of summary statistics. The concept of distance between nodes in a graph varies from the usual meaning, and refers to the number of arcs in the path connecting the two nodes. Connectivity is defined as the degree to which the network as a whole and individual nodes or subsets of nodes are connected. As a structural property, connectivity provides insight into the relative simplicity or complexity of a network.

An arc (i,j) in our graph represents all flights between departure airfield i and arrival airfield j. The routing graph is easily transformed into the connection or binary connectivity matrix B = [b_{ij}], where b_{ij} = 1 if arc (i,j) exists, and zero otherwise. To find all the minimum distances d_{ij} between node pairs i and j, matrix B is successively multiplied by itself (powered) until all b_{ij} are greater than zero. Where B shows the presence or absence of direct (single arc) connections between nodes, the non-zero elements of B^2 = BXB are the number of two-arc paths between any two nodes. The power n that results in an element changing from zero for the first time is the minimum distance between its associated node pair. The power n such that all shortest paths had been found is the diameter of the network, and reflects the distance between the two nodes (or several pairs) most "remote" from each other. With these basic elements, we can define the graph-theoretic measures in Table 2.16.

The Konig and accessibility indices evaluate individual node connectivity. Accessibility indicates how "reachable" the node is. The most central places in the network have the lowest Konig indices. (A simpler measure would be the sum of the columns of B, which would show relative direct, but not indirect, accessibility). We would expect the most frequently used patient destinations and the staging facility locations to be the most central.

The remaining indices in Table 2.16 measure network connectivity. The beta index measures linkage intensity,

TABLE 2.16

GRAPH-THEORETIC MEASURES

Name of Index	Computational formula
Accesibility Index	$A = \sum_{j} d_{ij}$
Konig Index	$K = max\ d_{ij}$
Beta Index	$\beta = \frac{a}{n}$
Gamma Index	$\gamma = \frac{2a}{n(n-1)}$
Cyclomatic Number	$\mu = a - n + g$
Alpha Index	$\alpha = \frac{a - n + g}{n(n-1)/2 - (n-1)}$
Dispersion Index	$D(G) = \sum_{i=1}^{n} \sum_{j=1}^{n} d_{ij}$

where

n = the number of nodes
d_{ij} = distance from node i to j
a = the number of arcs
g = the number of subgraphs

the average number of arcs per node. Gamma computes the level of connectivity (or sparsity) as the ratio of actual to the maximum possible number of arcs. Cyclomatic number is the level or degree of redundancy in a graph. The alpha index relates actual to maximum redundancy. Dispersion refers to the degree of network compactness as a function of the spatial arrangement of the nodes. Dividing the dispersion index by the maximum possible number of nodes gives average distance from any node to any other node.

TABLE 2.17

AEROMEDICAL ROUTING NETWORK STATISTICS

Number of arcs (a)	1405
Number of nodes (n)	215
Connectedness (γ)	.061
Cyclomatic number (μ)	1191
Redundancy (α)	.052
Linkage intensity (β)	6.53
Diamter	7
Dispersion (D(G))	132957
Average distance	2.88

TABLE 2.18

NODES WITH HIGH CONNECTIVITY

		Konig	Incidence			Number of
Base	Accessibility	Index	In	Out	Total	Accesses
BLV	373	4	78	92	170	367
SKF	413	4	53	49	102	218
ADW	417	4	57	55	112	186
BIX	451	4	40	30	70	160
SUU	496	4	25	25	50	84
BKF	527	4	18	24	42	106
SPS	512	4	21	17	38	45
SDF	516	4	12	14	26	29
ORF	501	5	21	15	36	55
LSF	547	4	16	17	33	43
IAB	544	4	11	12	23	21
HIF	540	4	6	7	13	29
VPS	551	4	9	14	23	37
TCM	574	4	9	10	19	37
SLN	567	4	7	9	16	50
NKX	583	4	13	15	28	46

Table 2.18 lists the most central, accessible, directly connected and most frequently transited places (shown as the total number of accesses, the sum of flights into and out of each node). The same nodes that are major patient destinations are also the most accessible in the route structure. Frequency counts of the Konig index for the 215 nodes were 48, 134, 32 and 1 for values 4 through 7 respectively. Only Scott had an accessibility index less than 400, and only 10 less than 500. The majority were in the range of 500-800; only five nodes exceeded 800. The highest was 998. There were 97 in the 501-600 range, 58 between 601 and 700, and 44 from 701 to 800.

Despite the fact that the structure is relatively weakly connected, without a significant degree of connectedness and redundancy, the maximum distance (diameter) is only seven, and only one node pair is that remote. The apparent reason is a relatively flat hierarchical structure, with a few central places connected to a large number of nodes by single arcs or very short paths. Further evidence of a highly branched hierarchy is given by other characteristics of the network.

For a number of reasons, path lengths are relatively short. First, we have already noted the tendency toward intraregional transfers among relatively close facilities. Secondly, the range of the aircraft is limited, such that many node pairs cannot possibly be directly connected. And third, this network shares a characteristic common in many transportation systems: since the central places are relatively far apart, long interconnecting route segments, even ones that carry high volumes, would tend to bypass stops where additional patients could be served, and schedulers are very reluctant to miss an opportunity to make an extra pickup or delivery.

To verify some of these assertions, we extracted a number of network attributes from the flight data. Figure 2.15 classifies route segments by flying time, frequency, and type of connection. The flying time histogram shows that the average flight duration was 1.14 hours, which corresponded to an average great circle distance of 339 nautical miles. Segment frequency is markedly skewed toward low values; more than one third of the 3663 flights were the only one or two over a given segment. Again, this probably reflects the impact of demand-responsive scheduling. Functionally, the highest proportion of flights were within regions, owing to the fact that 43 per cent of all transfers were intraregional. Interregional flight activity was substantially less. Flights between the central base and between staging bases were flown in numbers disproportionately high relative to the limited number of possible connections.

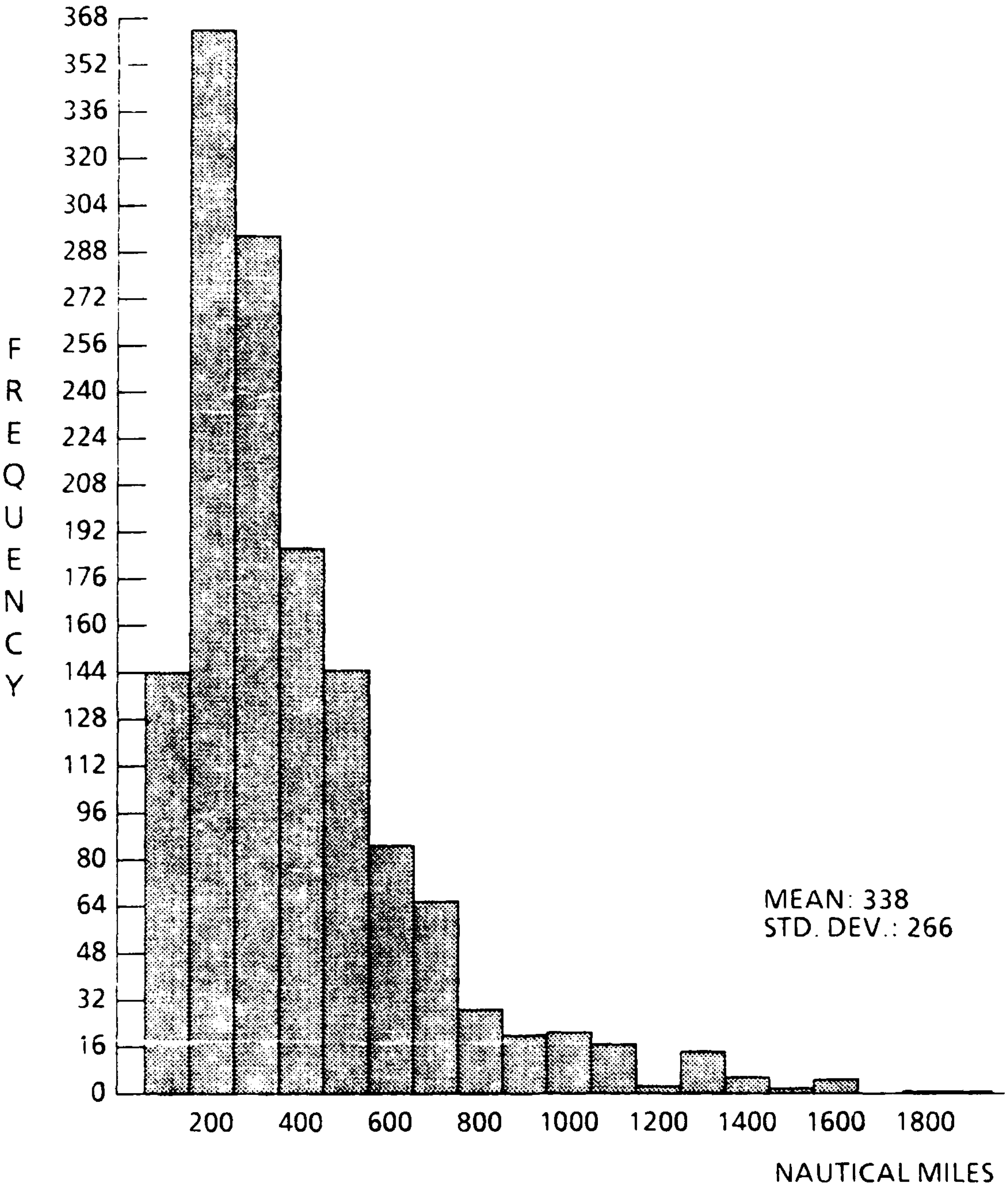


Figure 2.15a. Distribution of route segment distances.

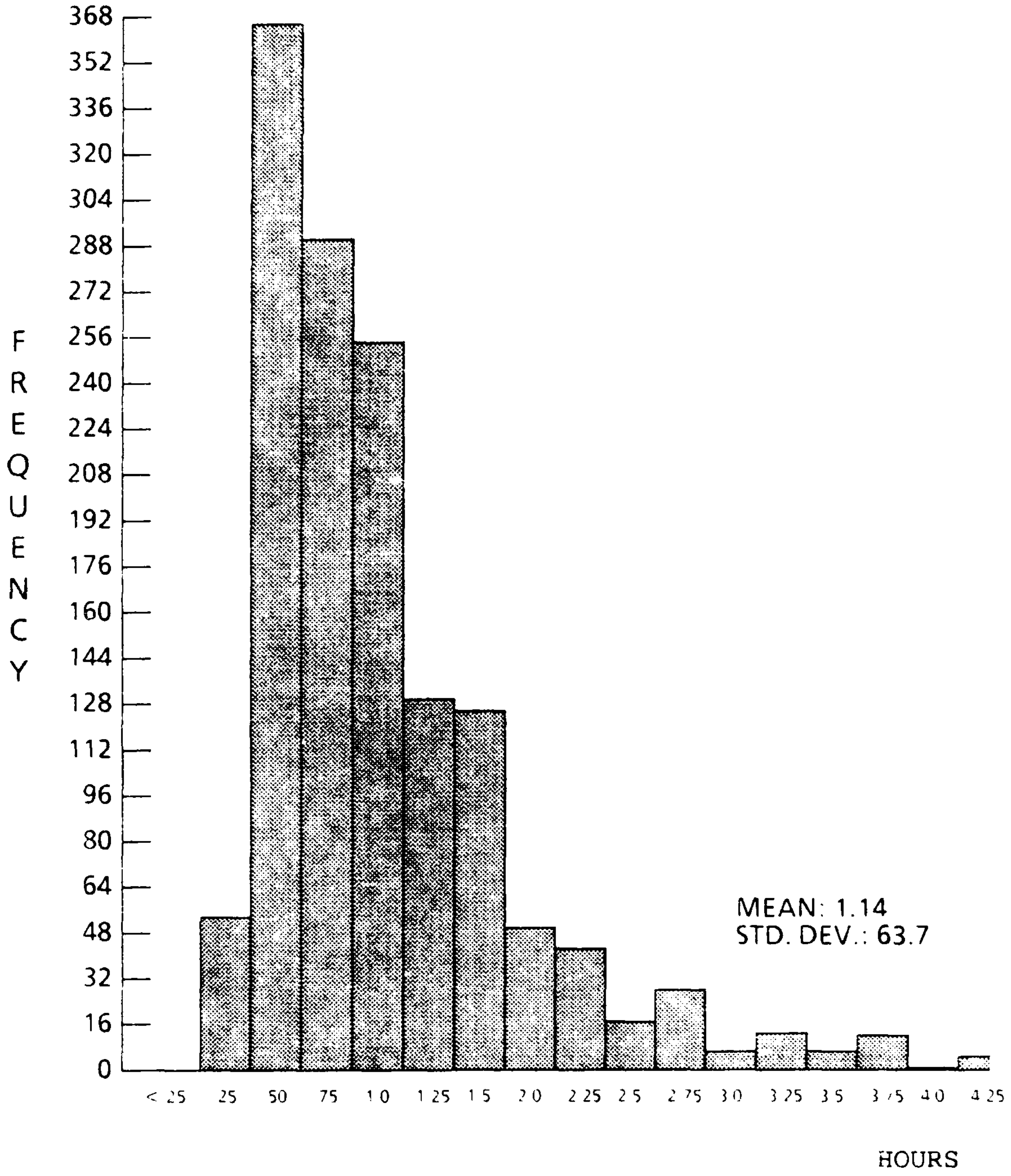


Figure 2.15b. Distribution of route segment flying times.

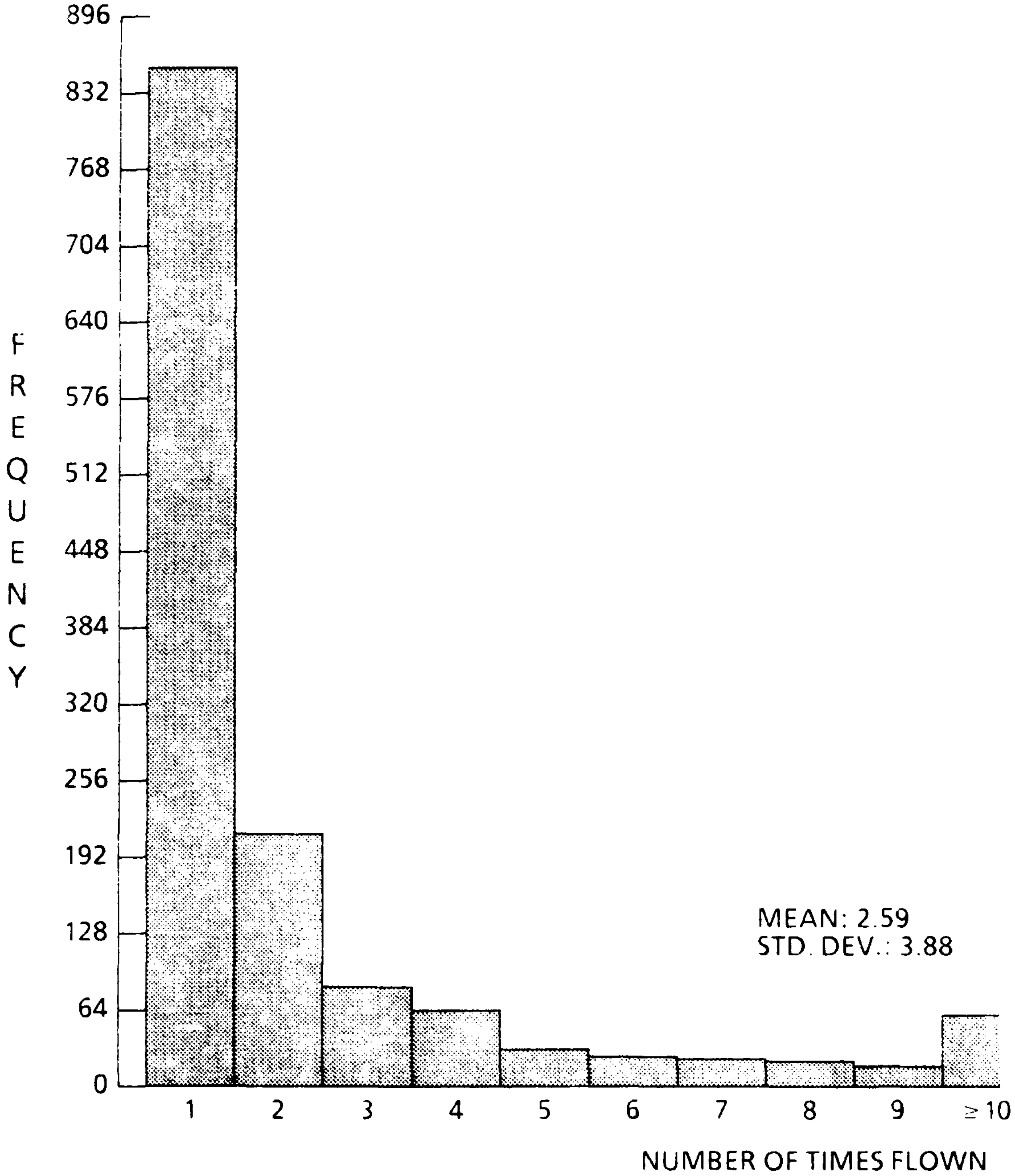
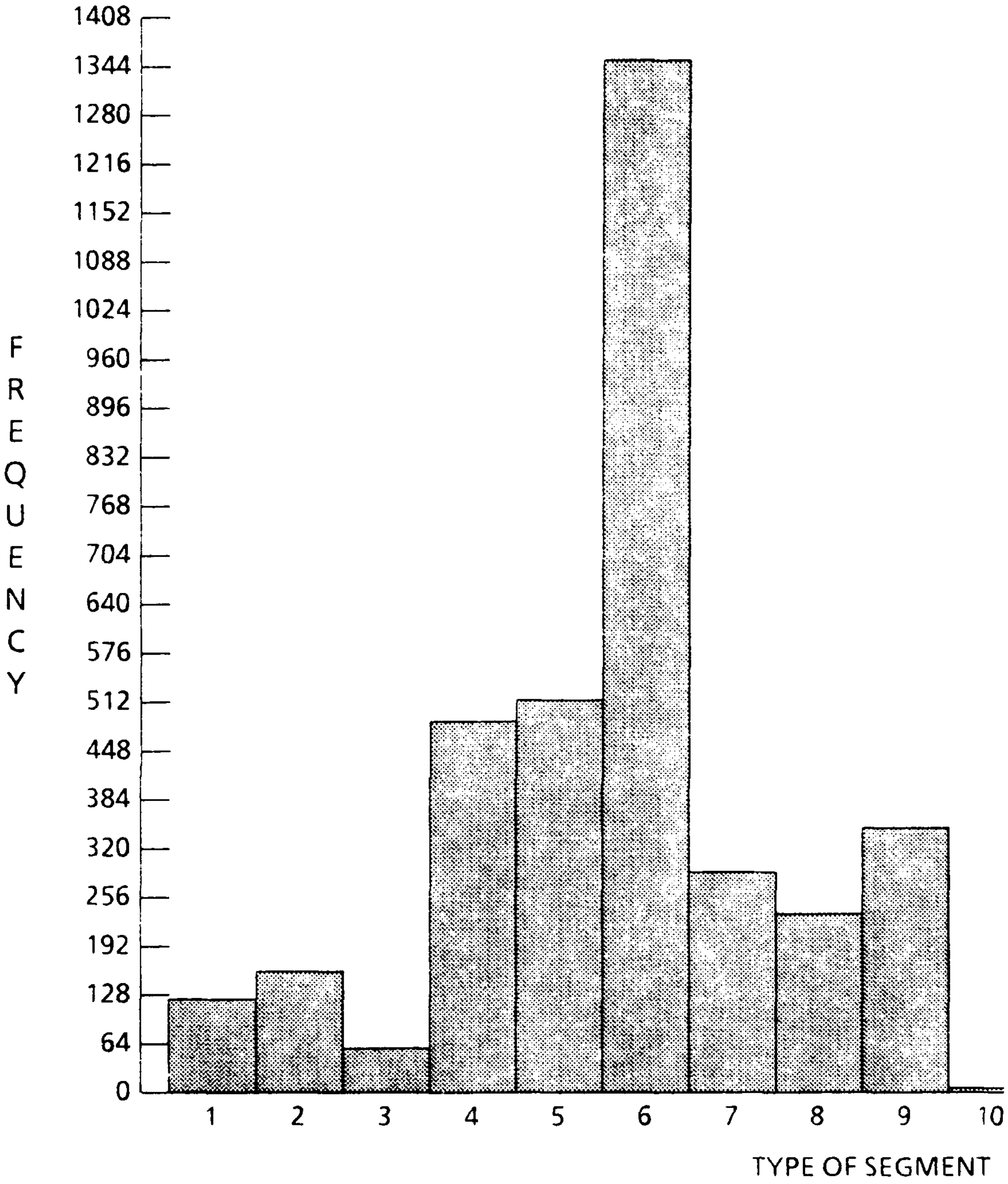


Figure 2.15c. Route segment frequency.



Type	Description	Type	Description
1	BLV --> ASF	4	ASF --> Non-ASF, Same Region
2	ASF --> BLV	5	Non-ASF --.>ASF, Same Region
3	ASF --> ASF	6	Non-ASF --> Non-ASF, Same Region
10	Urgent mission	7-9	Same as 4-6 except interregional

Figure 2.15d. Distribution of route segments by type.

This preliminary view of the route structure seems to indicate a route hierarchy involving the most central places connected by high-volume trunk routes, and other less important places served by medium-volume regional and interregional feeder structures. But, this accounts for only one-third of the total number of route segments flown. Patient flows indicated a concentration of low-volume nodes clustered around or between regional centers in collector/distributor subnetworks. To determine if there are corresponding routing substructures, we applied principal components and factor analysis to the binary connectivity matrix, techniques that are widely used to find partitioned subregional networks. [LOWE75] The techniques used are the same as those employed in patient flow analysis. The results are summarized in Table 2.19

The first factor produced high loadings on arcs eminating from the central base (Scott). In general, bases with very high numbers of arcs incident out of them induce high variation. In effect, Scott's associated region was much of the US insofar as many one-time flights, particularly urgent misions, orginated there. Factor loadings varied inversely with the incidence into the base connected with Scott. In other words, the high degreee of variance explained is attributable to the high degree of connectedness Scott has with a large number of bases.

TABLE 2.19

AEROMEDICAL ROUTE NETWORK FACTOR ANALYSIS RESULTS

		Per Cent
Total Variance	215.0000	100.00
Variance Explained by Each Basic Pattern		
1	23.3075	10.84
2	13.5435	6.30
3	14.0071	6.51
4	8.4183	3.92
5	7.1063	3.31
6	6.7725	3.15
7	6.5691	3.06
8	5.4825	2.55
9	6.1669	2.87
10	4.5016	2.09
Totals	95.8753	44.60

The second and third factors show similar relationships between Andrews and Kelly and bases to which they are directly connected, although virtually all the bases are in the respective regions served by the two. Two other bases in region 2 had factor scores one or more standard deviations above the mean, reflecting the presence of subordinate bases with relatively high connectedness, though substantially less than the regional centers.

This effect increases through the remaining factors, i.e., each factor includes one or two bases with very significant (by five or more standard deviations) factor scores and several others with scores one to four deviations away from the mean. In some instances a smaller pattern existed with factor scores varying in the opposite direction, but these were usually better accounted for in other factors. If the interconnections among bases with high factor scores are recorded in a connectivity matrix, the row sums and factor scores display a very discernible relationship. This indicates that the smaller subnetworks selected by lower order factors are fairly highly interconnected, with typical sparsities of .35 to .50. These subnetworks tend to be very tightly concentrated geographically. (See Figure 2.16) This should be expected, since we have already seen that relatively short, intraregional route segments are the most frequently flown. Other subnetworks extracted by each factor were:

Factor	Subnetwork Area	Principal Base(s)
4	Southwest (Southern Region 4)	Travis (SUU)
5	Mississippi-Florida	Biloxi (BIX)
6	Region 3 and Kansas	Buckley (BKF)
7	Southeast (Region 2)	Augusta, GA (AGS) Fort Benning, GA (LSF)
8	Region 6	Wright-Patterson (FFO)
9	New York-New England (Region 1)	Limestone AFB (LIZ) Plattsburgh AFB (PBG)
10	Pacific Northwest	McChord AFB (TCM)

The temporal pattern of flights was derived from the flight segment data set, which contained some 3600 segments. In several years, the basic weekly and monthly schedules have not changed, despite the claim made by some system managers that demand not only had increased, but

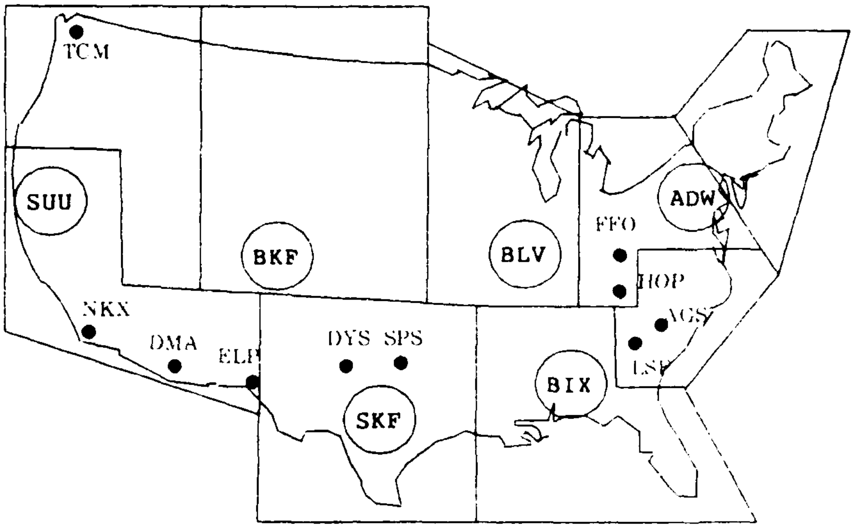


Figure 2.16. Regional Subnetworks.

that the basic patterns had changed substantially. The most significant sources of variation between scheduled and actual routes were changes in patient condition and weather, and the need to reroute routine missions to handle roughly 60 per cent of all urgent cases. [DEPA78b] Figure 2.17 shows the typical patterns and numbers of flights each day. The 490 routine missions ranged from three to ten segments (most flights were seven or eight) and up to 11.0 flying hours and 16.0 total crew duty hours in length. 56 urgent missions (not shown) ranged from two to six segments and up to 11.3 hours in length; most served one or two urgent patient movements that could not be handled by routine mission rerouting, and departed and returned to Scott AFB. Because many urgent cases are handled by preempting routine missions, the daily schedule can be considerably

DAY	AIRCRAFT SCHEDULED
Monday	6
Tuesday	7
Wednesday	6
Thursday	6
Friday	7
Saturday	6
Sunday	4
Total	42

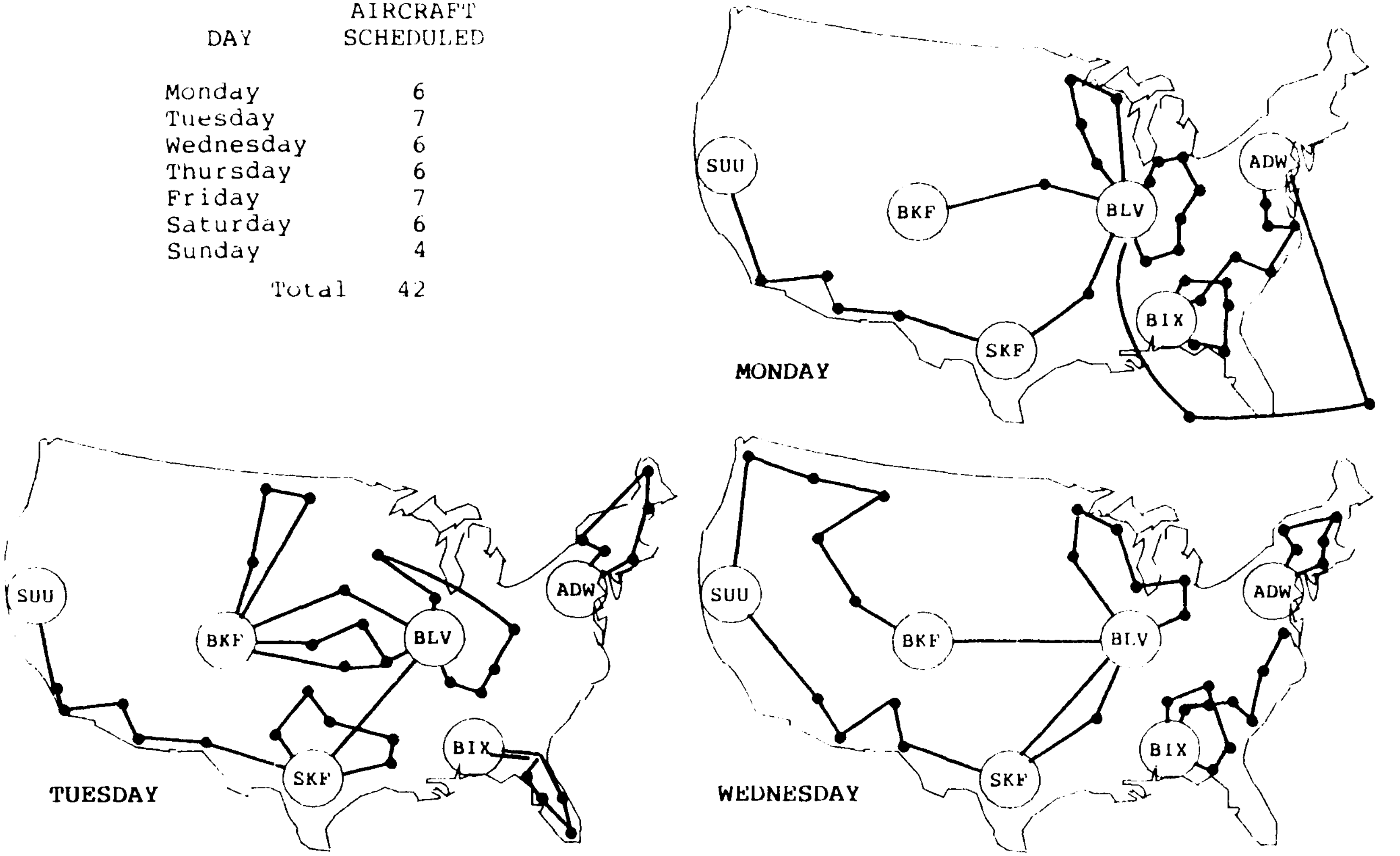
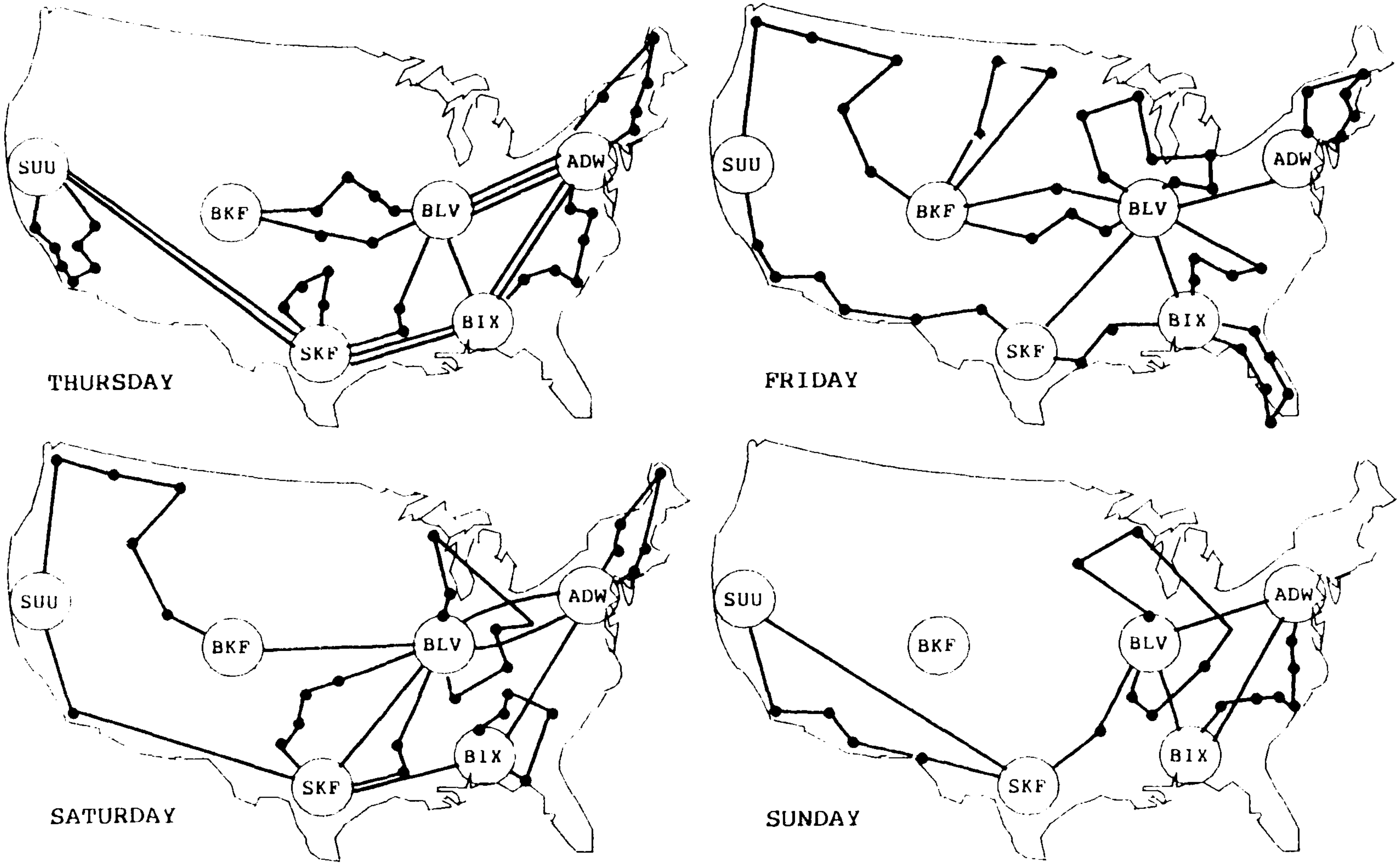


Figure 2.17. Weekly Flight Schedule.



THURSDAY

FRIDAY

SATURDAY

SUNDAY

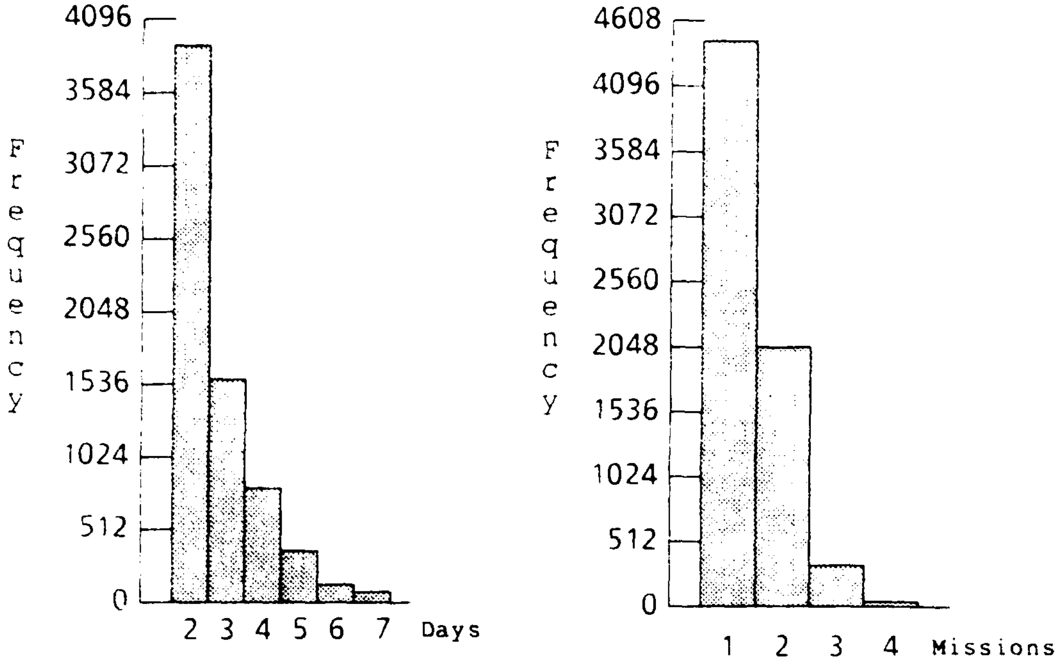
Figure 2.17. Weekly Flight Schedule (Continued).

disrupted, forcing routine patients to make unscheduled overnight stops.

By virtue of the fact that the system regularly served some 50-60 major origination and destination points, and also had to stop at another 150 over a ninety day period, many of those only once, the most critical problems that emerge are the choice and sequencing of stops, which are limited to eight per mission, and designating the starting and ending points for each mission, since these directly affect preceding and following days. As the weekly schedule shows fairly clearly, some regions do not receive any service on three or four days. The slack available on Sunday is not readily useable, given the pattern of weekday demand concentration we saw in the flow data. And, given the already relatively short flight durations, reducing individual segment distances is not as important in this problem as in others where cost minimization is the para mount goal. Such reductions will not greatly increase the number of stops that can be made, which is driven to a significant extent by ground time at each stop (20-40 minutes) and the two-hour preflight period, both of which account for about one-half of the 16-hour maximum crew duty day of an eight-stop mission. Recalling our discussion of the choice of patient destinations, we can see that that could be an area where significant improvements could be realized, by reducing the number of unnecessary stops.

2.3.4 System Performance. Attempting to characterize individual and aggregate patient trips is difficult because the movement data contained only final movement details. No record was available to show the intended movements and all changes. Because of reroutings to serve higher precedence patients, routine patients were frequently subjected to unplanned overnight delays. But the data does reveal some general characteristics of actual trips, such as those in Figure 2.20. Total time in-transit includes all time elapsed from the first day the patient could have been moved. The number of different missions gives a general idea of how many overnight delays patients experienced, either because of unplanned routing changes or because their destinations were not served by the same mission that picked them up. System managers, of course, want to reduce both in-transit time and the number of enroute delays patients have to endure.

We also examined how well the system met the criteria for timely pickup and delivery. Routine patients were picked up within 48 hours in 5753 of 6907 cases; only 446 were not picked up within the 72-hour limit. 10 priority patients were not picked up within one day, and all urgent cases were provided same-day service. 5901 routine patients were delivered either the same day or the next day after being picked up, and only 118 were not delivered within the 72-hour limit. One priority patient out of 170



Total Time In-transit **Number of Missions Flown**

Figure 2.18. Patient trip characteristic distributions.

was delivered in more than one day, and all urgents were picked up and delivered in one day. System managers said that if complete delivery could not be provided on the same mission, a delay at the pickup point was usually preferable to enroute stops. However, movement toward the destination, particularly in movements between distant points, was better facilitated if patients were moved to intermediate points closer to the destination.

It is much more difficult to assess service quality, or even specify measurable attributes of quality. As we discussed in the section on performance measures, the number of stops any one patient makes is a function of his own transfer points and those of all others on-board the same airplane. Certainly, the most direct flights are preferable to the individual patient, but since the total number of stops is a significant limit, directness has to be balanced. If we look for routings that will allow additional stops, we might improve the overnight delay problem, but we will reduce the quality of service in other respects, for example, by reducing directness.

Further research is needed to determine appropriate measures of performance. If patient diagnostic codes are available, techniques such as logit analysis could be applied to quantify such things as destination assignment tendencies, and some of the underlying biases we asserted in moving patients between hospitals operated by the same service branch. The benefits and costs of travel to individual patients are not well understood, and are confounded by such things as the lack of population, workload and care cost data that allow other alternatives to travel to be adequately assessed.

2.3.5 Formulation of the Problem. From our discussions of the general characteristics of the DOD Military Health Services System and the nature of the aeromedical transportation problem, it should be evident that attempts to improve the transportation subsystem must be made in the context of improving the whole DOD medical program. Otherwise, to simply concentrate on making it more "efficient" may lead us to do, albeit very precisely, the wrong thing.

We should begin with an idea of what we want to accomplish. Even this is difficult, because of the current organizational structure, institutional planning and decision practices, operating rules, and perhaps most critically, the lack of suitable planning information and the mechanisms to collect, process and distribute that information for decision making.

If these problems were resolved, our ideal for the DOD medical system would be that depicted by the diagram in Figure 2.19. The hierarchical structure is intended to

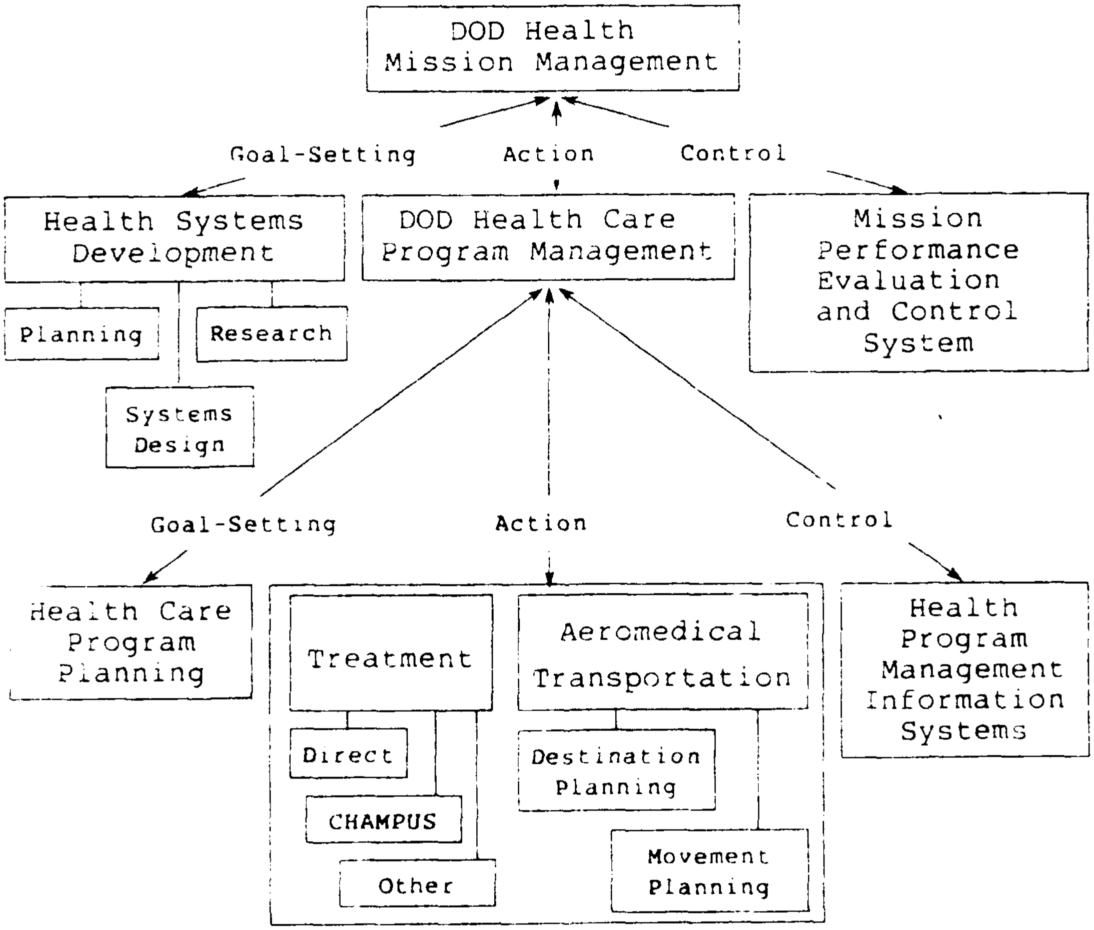

DOD Health Mission Management
Goal-Setting
Action
Control
Health Systems Development
DOD Health Care Program Management
Mission Performance Evaluation and Control System
Planning
Research
Systems Design
Goal-Setting
Action
Control
Health Care Program Planning
Treatment
Aeromedical Transportation
Health Program Management Information Systems
Direct
CHAMPUS
Other
Destination Planning
Movement Planning

Figure 2.19. An ideal DOD health management system.

convey two principal concepts: a decentralized allocation of responsibility for planning, activating and controlling missions and individual mission programs; and a coordination mechanism by which the actions of all system participants are directed toward enhancing whole system performance. Current practices of isolating and confining decision making, enforcing strict hierarchical, preemptive or non-reviewable, sequential decision processes would be replaced with interactive processes of the type we recommended for destination and movement planning. The current system seems to be particularly lacking in the ability to utilize data and information on past resource allocations and system performance to find ways to improve the future use of resources.

With programs underway to create such system elements as the Resource Analysis and Planning System for the Assistant Secretary of Defense for Health Affairs office [STCL82], DOD-wide enrollment of all eligible beneficiaries [DOYL82], a uniform set of cost accounts for all health care activities [FRAG82], and facility treatment cachement area identification, many of the informational deficiencies will be eliminated. Rather than wait for these efforts to be completed, we propose a planning model that can support the health care system that DOD eventually adopts.

We will develop a transportation resource allocation mechanism that will provide service no worse than is currently provided, but will identify ways to better use resources to improve and provide increased opportunities for patient movement. At a minimum, the transportation planning system should incorporate the salient characteristics of the aeromedical system we have identified: patient movemenr response criteria for pickup and delivery; patient flows that vary structurally and over time; operating rules and the lack of aircraft that together do not allow all demand points points to be serviced each day; the need to conserve fuel and other resources by finding routes that reduce travel time, but not patient service; and the need to respond to special patient requirements. Most importantly, the model should find ways to transfer patients as expeditiously as possible.

For the long run, we should design the mechanism so that it can be embedded in the global DOD health management system, where it would have access to data on costs, facility capacities and capabilities, patient demands, etc., with which to make more comprehensive cost-benefit decisions. It should also be capable of providing answers to larger concerns, such as what tradeoffs can be achieved between changing hospital capacities and capabilities and increasing or decreasing the need for transportation, in terms of client benefit.

ENDNOTES

1. In the absence of standard terminology we will use aeromedical transportation to refer to the movement of patients by aircraft to or between medical facilities. The term aeromedical is formed from the root words aerial and medical. In emergencies or wartime, the term aeromedical evacuation is often subsitutued. The military acronym 'CONUS' refers to the continental United States, and domestic to the aeromedical system within the US.

2. The cost of equipping a helicopter ambulance is about $500,000, and about the same amount is needed annually to staff and operate it. [HELI83]

3. C-141B is the US Air Force designation for the Lockheed Starlifter transport. The C-9A Nightingale is the specially modified McDonnell Douglas DC-9 commercial transport used exclusively for aeromedical transportation.

4. As recent studies indicate, the exact population of eligible beneficiaries is not known. Recent efforts to establish an identification system may produce the first accurate estimate by 1984. Not only is the exact size of the population unknown; beneficiary location with respect to medical facilities has only recently been studied. [DOYL82]

5. Because active duty personnel are represented in both objectives, it might appear that the first objective is

subsumed by the second. However, in the first the beneficiary of goal attainment is the organization, while the individual benefits from the second.

6. Evacuation points are those airfields in the US to which the evacuated casualties are flown from a war theater on temporarily converted cargo airplanes.

7. We base our claim of stability on statistics on the level of aeromedical services from 1973-1983, which show a fixed number of hours flown each year, and only a modest increase (approximately 4 per cent per year) in patients carried, reflecting reductions in clinical staffs and facilities and in entitlement programs. [JONE82]

8. This deficiency holds for a European conflict, as a recent planning conference established, and is apparently true for potential conflicts in other parts of the world. [DEPA83b] Even during the Vietnam War, patient care and transportation requirements did not severely tax the Military Health Services System.

9. Again, lacking data on casualty estimates and bed availability, the extent of the redistribution problem is unknown but presumed to be large.

10. A mission is the set of flight segments covered by one aircraft in one day.

11. The C-141B can carry from 107 to 167 patients, depending upon the amount of special medical equipment used. The C-9A maximimum load is 44 patients.

12. Of the total fleet of eleven, one or more aircraft may not be available due to maintenance and other problems.

13. These factors should be considered assertions to be verified later in our analysis of patient movement data.

14. DOD utilizes two different forms of organization. The traditional form assigns responsibility to train and equip naval, land and air forces and construct and maintain facilities to three separate military services. Forces are actually employed under a mission-oriented scheme, with components drawn from any of the services, under a single commander designated by the Joint Chiefs of Staff. Under the dual organizational arrangement, medical and operational units assigned to the same administrative organization might not be employed together under the mission-oriented scheme.

15. In extenuating circumstances, such as accidents, soldiers may be treated at civilian hospitals. Military hospitals may also contract with nearby civilian hospitals for services they cannot provide.

16. We did not consult the obvious source, the doctors themselves. With over 8000 patients moved, this was not

practical. Doctor-patient confidentiality restrictions also ruled against this.

17. Normal bed capacity is the capacity for normal peacetime use of space.

18. Operating beds are those set up and staffed for care of a patient.

19. Occupancy rate is the number of operating beds divided into the average number of daily occupied beds.

20. The Military Airlift Command provides other transportation services that are paid for by users. The Airlift Services Industrial Fund is a working capital account that introduces a quasi-market pricing mechanism into DOD transportation planning. MAC charges each user a tariff for services, and it funds the bulk of its operations with the revenues. It cannot charge tariffs out of line with other alternatives available to users, as they have the option of utilizing a less expensive alternative to air transportation. By DOD policy, all patient transfers in excess of 100 miles must be made by air. [JONE82]

21. Flying time is the elapsed time from takeoff until the aircraft is parked at a destination. It is a ubiquitous unit of analysis in Air Force planning commonly considered an output, although from a budget standpoint, it represents the expenditure of inputs.

22. That price is an average of all forecast fuel purchases for one year. Bases that pay actual fuel costs above the uniform price are subsidized by transfers from bases that pay less. This scheme ensures that bases do not show disproportionately higher costs than other bases for their flying operations. It also virtually prohibits any economic fueling strategies based upon price, since it was designed primarily for non-transient aircraft operations. Wing policy does direct crews not to purchase from non-Air Force sources, since actual costs, including refueling charges, ramp fees, and profit, must be paid.

23. Ordinarily, the route covered by one aircraft in one day constitutes a <u>mission</u>. Exceptions include instances where an emergency requires an aircraft to be used for a second mission, or when an aircraft maintenance problem requires a second aircraft to complete the mission.

24. As an example, a patient with a severe pulmonary problem may require a constant aircraft cabin air pressure that can only be maintained at low flight altitude. It is the increased fuel consumption at that altitude, and not the medical condition of the patient, that may impose a route length restriction in the routing problem.

25. A patient's records show only his enplaning and deplaning airports. We could not always determine conclusively the aircraft route, particularly when the route transited the same airport more than once.

CHAPTER III

MODELS FOR DOD DOMESTIC AEROMEDICAL TRANSPORTATION PLANNING

For every problem there is one solution that is simple, neat, and wrong. H.L. Mencken

3.1 Introduction. In the last chapter we observed that the ability to serve patients was limited primarily by the scarcity of two resources, aircraft and crews. Fixed aircraft fleet size, the desire to use the limited life of each aircraft sparingly to preserve wartime capability, and aircraft maintenance policies dictate the number of aircraft available each day, which governs the number of routes that can be flown. Safety and patient concerns limit flight and medical crew duty time and the number of stops allowed in a duty period, which restricts the number of segments in and duration of each route. As a result, on most weekdays, the number of patient origins and destinations exceeds the maximum total stops allowable.

Aeromedical transportation planning, then, involves more than daily aircraft routing and crew scheduling; it must also resolve the imbalance between movement demand and resource availability through an appropriate allocation of resources over longer time periods. And, perhaps most importantly, the planning process must take place within the institutional context and operating environment described in the last chapter.

The objective of this chapter is to propose a design for a planning system that will help planners and decision makers resolve these problems. We begin by outlining an approach that will examine a series of increasingly difficult aeromedical decision problems that unfold from the relatively simple problem of finding daily routes into the very difficult problem of resolving imbalances between patient movement demand and resource availability while observing significant operating restrictions. We then discuss the particular methodology to be used, a variant of linear programming that explicitly reveals how resource use and client service are coordinated and controlled as the planning model finds improved planning solutions. The chapter concludes with a proposed resource-directive LP model, and a review of related models in the literature.

3.2 Approach. In our discussion of planning systems in the last chapter, we described how Mason classified planning systems by the design of their information and decision-making subsystems. That scheme creates an interesting dilemna: which of the two subsystems do we design first? Simon observes that in the past the design process

> most often began with asking what could be done with the information that was already there, not with asking what decisions were being made, and what information would be helpful in making them. [SIMO77,p. 126]

In many instances, information systems designs were based on an historic or combination user/historic model. [AMER72,p.319] Designers asked decision makers what data they used; where it originated and how it was being captured, processed, communicated, retained and transformed into information for decision. But, as Mason points out,

> All too frequently, however, the approach used by designers has been limited to a study of existing forms, files, reports and procedures and an effort to determine ways in which they might be simplified, expanded, integrated and improved. Sometimes many useful and economical results are realized from this kind of study; but, they are generally in the area of increased efficiency of data flows or reduction in clerical staff, *not* in improved quality of decisions made. [MASO75,p. 3]

Ackoff [ACKO67] argues that using the approach *can have* far more serious consequences. When designers ask managers what information they need, they assume (1) managers fully understand the types of decisions they make and have adequate models of each type, and (2), they really want the data they claim to need to make those decisions. [ACKO67] Although it is unlikely that both conditions are always satisfied completely, managers probably have adequate conceptions of at least some of the decisions they make, and their demands for data may adequately support their needs. But the critical problem is this: less than full understanding of the nature of the decision problem causes managers to consider more ariables and hence ask for "everything." The designer, even less familiar with the manager's decisions, incorporates even more variables, generating more than "everything."

The resulting system typically deluges the manager with an "overabundance of irrelevant data," [ACKO67, p.148] and fails to perform the two most important functions of an information system, data filtering and condensing:

> The designers of the[se] system[s] had not learned the first lesson of living in an information-rich world: that a major task of an efffective information system is to filter information, not to proliferate it. [SIMO77,p.294]

The vast literature on information systems is a litany to other shortcomings of the user/historic approach. Because one of the most common data sources managers rely on is an accounting system, many systems are predominantly transactions-oriented, in the accounting sense that, if something measurable does not happen, the system does not record it. Such systems react passively to the environment because they do not "know" anything until and unless an event has taken place and the fact recorded, rather than actively attempting to forecast and anticipate significant environmental conditions. The approach also usually fails to record a suitable history of resource use, including lost opportunities, because no audit trail is kept of planned events that did not actually occur. [CHUR67] This is a problem we noted earlier in the current aeromedical information system.

Also, using the current organization as a model avoids or precludes considering better alternatives. And, the time delay between finalizing the design and implementing the information system induces rigidity in organizational structure and functioning by preserving outmoded practices and procedures. [HEDB76] Finally, basing a design on the preferences of current participants and their practices often produces systems that are incompatible with different user psychological types and other managerial styles. [MASO73] In Wildavsky's view, the approach has not merely failed, it has never suceeded. [WILD75,WILD76]

The underlying problem, Ackoff suggests, is that:

> One cannot specify what information is required for decision making until an explanatory model of the decision process and the system involved has been constructed and tested. [ACKO67,p.150]

Mason [MASO75,p.2] agrees; of two approaches to planning system design, studying and refining data flows, versus analyzing management decision problems and conceptualizing and modeling their essential elements, the latter is better suited for planning system development. Designers should first determine what decisions must be made and how, by means of a model with the components of the decision process shown in Figure 2.1. Then, the parameters and variables in the decision model serve to specify what data is required, where it originates, and how it should captured, processed, distributed, and retained; and how it can be most usefully transformed into information for decisionmaking. We state as a major premise of this paper that formulating a decision model first is fundamental to planning systems design.

The major methodological problem, then, is to decide how to construct a decision model for DOD aeromedical transportation planning. The aeromedical problem as we described it in the last chapter includes a large number of potential decision problems. We are primarily interested here in solving routing-related problems, finding sequences of visits to all patient service points, subject to constraints on resource availability and operating limitations, that maximize or minimize some objective measure(s) of performance.

Bodin and Golden [BODI81] provide a very useful way to categorize transportation problems according to the set of characteristics common to most problems they have analyzed (Table 3.1). Essential aeromed problem characteristics discussed in Chapter II are denoted in Table 3.1 with asterisks. By varying our assumptions, we create a series of increasingly more complex models in terms of these characteristics (Table 3.2).

TABLE 3.1

AEROMEDICAL TRANSPORTATION PROBLEM CHARACTERISTICS

	Characteristic		Possibilities
A.	Number of depots	1.	One depot
		*2.	More than one depot
B.	Fleet size	1.	One aircraft
		*2.	More than one aircraft
C.	Type of fleet	*1.	Homogeneous aircraft
		2.	Heterogeneous aircraft
D.	Nature of demands	*1.	Deterministic
		2.	Stochastic
		3.	Static
		*4.	Dynamic
E.	Demand location	*1.	At nodes
		2.	On arcs
		3.	Mixed
F.	Underlying network	1.	Undirected
		*2.	Directed
		3.	Mixed
G.	Aircraft capacity limits	*1.	Imposed-identical aircraft
		2.	Imposed-not all the same
		3.	Not imposed
H.	Maximum route times	*1.	Imposed-identical limits
		2.	Imposed-not all the same
		3.	Not imposed
I.	Costs	*1.	Variable routing costs
		2.	Fixed operating and/or acquisition costs
		3.	Variable and fixed costs
J.	Service operations	1.	Pickup only
		2.	Delivery only
		*3.	Mixed
K.	Service stipulations	1.	Serve all customers
		*2.	Service deferment or partial service allowed
M.	Problem-dependent constraints	*	

Source: [BODI81,p.98]

Version I is the single aircraft routing problem, in which the object is to find the route or routes that provide complete service to a set of customers. The essential feature of patients having both origins and destinations is introduced in Version II. This changes the problem from one that is essentially the classic traveling salesman problem to one that is considerably more difficult and requires a new algorithm to solve. In addition to finding a route initially, the procedure can also be used to quickly revise the routing if new stops must be added or previously included stops changed or deleted. These situations occur frequently, for example, when missions in progress must be revised to handle urgent cases. Chapter IV will discuss the first two versions.

Version III introduces the problems of designing routes for a fleet of aircraft and operating from multiple depots. (A depot is the more generic term in the literature for what were earlier called aeromedical staging facilities). In Version IV, we introduce the complications caused by shortages of aircraft and crews, the effects of geographic regionalization, and the need to cycle aircraft through the central base periodically. Versions III and IV will be covered in Chapters V and VI respectively. Note that characteristics C, F, and G will remain constant in each of the model versions.

TABLE 3.2

AEROMEDICAL PLANNING MODEL DEVELOPMENT

Version	Characteristics	
I	A(1)	One depot
	B(1)	One aircraft
	D(1),(3)	Deterministic, static demand
	E(1)	Demand located at nodes
	F(2)	Directed network
	G(3)	No aircraft capacity limit
	H(3)	No route limit imposed
	I(1)	Variable routing costs
	J(2)	Delivery only (no precedence)
	K(1)	Service all clients
II		Same as Version I except:
	J(3)	Mixed service operations (precedence relationships present)
III		Same as Version II except:
	A(2)	More than one depot
	B(2), C(1)	More than one (identical) aircraft
IV		Same as Version III except:
	D(4)	Dynamic demand
	H(1)	Maximum route limits (time)
	K(2)	Service deferral and partial service allowed
	M	Problem-dependent constraints: Originating and final depot for each route are decision variables

In the following section, we will introduce the linear programming methodology we will use in the planning model. We will emphasize the properties of the LP model that will allow us to model time periods (e.g., days of the week) and regional organization, and decompose very large problems using the resource-directive LP decomposition technique that explicitly models the initial assignment and iterative reallocation of resources to better serve patients.

3.3 Methodology. Many organizational planning problems can be usefully represented as a mathematical program:

(R) Find $\underline{x}_k$, k=1,2,...,K, in order to

$$\text{Maximize } Z = \sum_{k=1}^{K} f_k(\underline{x}_k) \tag{3.1}$$

$$\text{Subject to: } \sum_{k=1}^{K} g_k(\underline{x}_k) \le \underline{b} \tag{3.2}$$

$$\underline{x}_k \in X_k \qquad k = 1,2,\ldots,K \tag{3.3}$$

$$\underline{x}_k \ge 0$$

where:

$\underline{x}_k$ (k=1,2,...,K) is a vector of activity levels with dimension n_k,

$f_k(\underline{x}_k)$ (k=1,2,...,K) is a real, scalar-valued function with argument $\underline{x}_k$,

$g_k(\underline{x}_k)$ is a real, vector-valued function associating resource utilization with activity level $\underline{x}_k$,

$\underline{b}$ is a vector of resources available to the organization with dimension m, and

X_k is the feasible region to which allowable vectors $\underline{x}_k$ are constrained.

The use of a singular, scalar-valued, multiple argument objective function, $f(\underline{x})$, is very common in the literature. The constraints, $g(\underline{x})$, are interpreted to be the organization's technology, its environmental restrictions, and other goals and stipulations not specified in the objective. Shown as inequalities, equations (3.2) have as their right hand side the resources available to the production process. These may also be interpreted as production targets or resource budgets. [CHAR61,HAAS68] Since a system of linear inequalities of the type shown in (3.2) cannot be solved directly, linear equalities must be created by adding slack variables, which measure the extent of resource underutilization or the underachievement of targets. [CYER63,pp.36-38] In the case of greater-than-or-equal-to constraints, surplus variables are used to create equalities that can measure overachievement and utilization of resources above some stipulated minimum. To both greater-than and equality constraints, artificial variables must be added that have no economic interpretation, in order to obtain initial solutions. In addition to the non-negativity constraint often imposed on the decision variables, $\underline{x}$ may be required to be continuous, or integer, or even both over different stipulated domains.

In the following discussion, we assume that functions (3.1)-(3.2) are linear, such that (R) becomes:

(R') Find $\underline{x}_k$, k=1,2,...,K, in order to

$$\text{Maximize } Z = \sum_{k=1}^{K} \underline{c}_k \underline{x}_k \qquad (3.4)$$

$$\text{Subject to} \quad \sum_{k=1}^{K} A_k \underline{x}_k \le \underline{b} \qquad (3.5)$$

$$B_k \underline{x}_k = \underline{d}_k \qquad k = 1,\ldots,K; \qquad (3.6)$$

$$\underline{x}_k \ge 0.$$

If constraint (3.6) is not present, (R') is often called the canonical primal form of a linear program (LP), which we discussed in the last chapter. Methods for solving (R') when the $\underline{x}_k$ are continous are widely known. In our planning model, decision variables which represent the decision to travel between two points must be integer. Ordinarily, an integer linear program (ILP) of even moderate size cannot be solved. However, some ILP problems can be solved because of special structural properties. Our formulations of Versions I and II of the aeromedical planning model, for example, are an extension of one ILP special case of (R') commonly called the assignment model. This allows us to use special techniques that solve the assignment problem efficiently, within a branch and bound framework that can handle additional constraints.

When constraint (3.6) is present, (R') is a linear program with block angular structure, in which the non-zero coefficients in the constraints are clustered in submatrices along the diagonal of the constraint matrix. Constraints of this type can occur when portions of an overall problem are nearly separable. In an organizational context, this might be due to geographic, temporal or functional specialization, or a combination of these. An extensive literature discusses models of the type shown in Figure 3.1(a), where the near separability of divisional problems creates a decentralized structure. Block

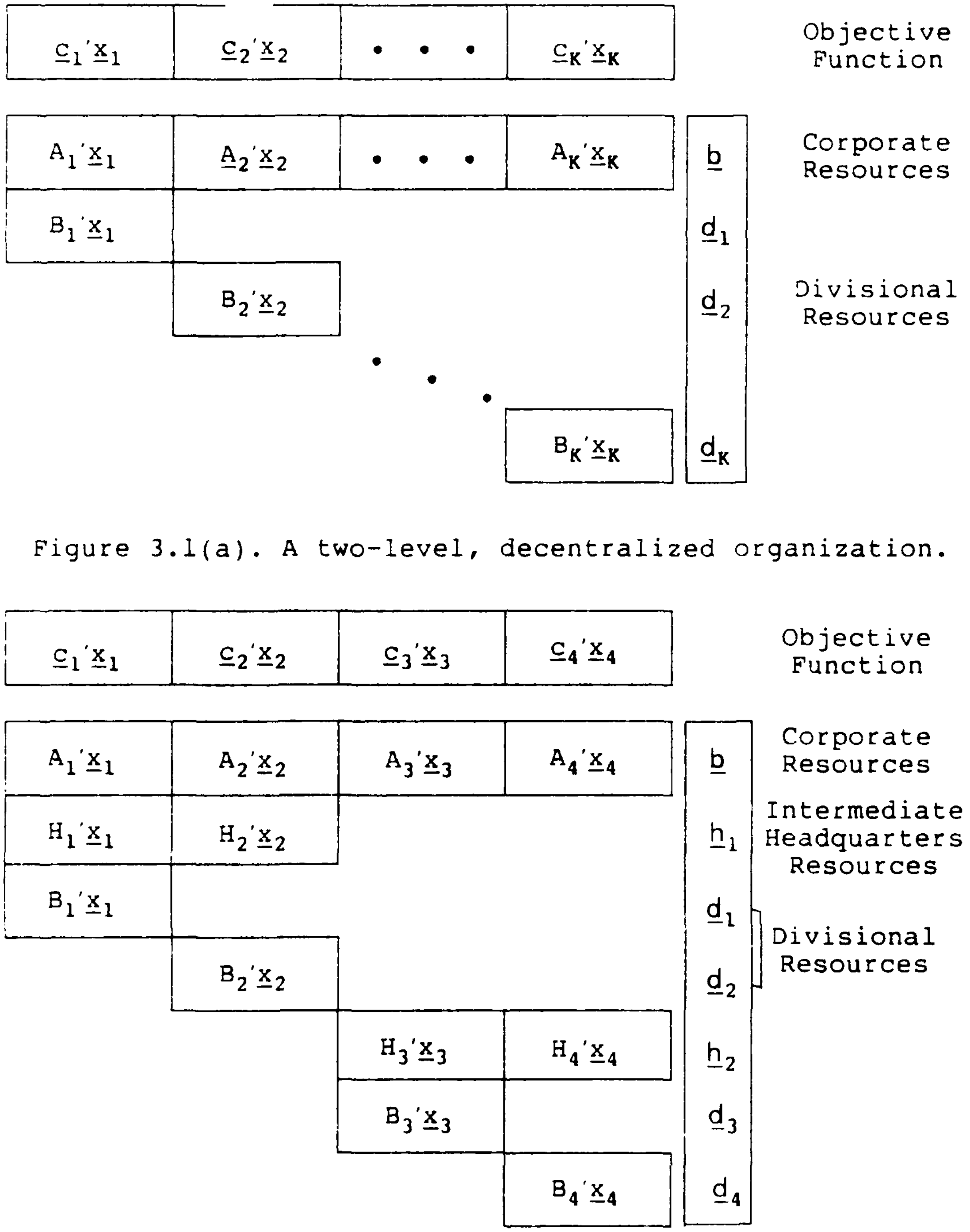


Figure 3.1(a). A two-level, decentralized organization.

Figure 3.1(b). A three-level, decentralized organization.

Figure 3.1. LP-based models of multi-level, decentralized organizations.

angularity can also be present within each major block, $B_k\underline{x}_k$, as in Figure 3.1(b). (R') can be interpreted as a model of resource allocation in a multi-level organization. Figure 3.1(b), for example, could represent a three-level aeromedical organization with a central management at the (wing) headquarters, geographic regional divisions, and patient transportation administrators located at the medical treatment facilities in each region. The objective (3.4) might be to minimize patient enroute delays, subject to two types of constraints: organization-wide limits (3.5) on resources required by more than one region, such as aircraft and aircrews, and regional resource limits (3.6) such as the maximum number of staging facility beds available. Resource vectors $\underline{b}$ and $\underline{d}_k$ are commonly referred to as <u>global</u> and <u>divisional</u> resources respectively.

An organizational LP model with block angular structure implies assumptions besides those discussed in Chapter 2.

1. The linear objective implies that the organization is cooperative. [FREE73] That is, although lower levels may compete for global resources, their objective attainments are additive. The model is not restricted to single objectives; Kornbluth [KORN74] discusses the multiple objective LP case.

2. Divisional resources are not transferable, or at least such transfers are not considered part of the allocation problem.[1]

3. The divisional problems are assumed to be independent, in that only activity levels $\underline{x}_k$ enter as arguments in g_k (i.e., the blocks $A_k\underline{x}_k$ and B_kx_k); there are no <u>externalities</u>.[2]

The problem of optimal resource allocation is usually not static, in the sense that maximal contribution from available resources is usually achieved over time. As we will show in Chapter 6, the daily demand for patient movement can exceed system capability, such that the time period (day or days) in which service is provided is part of the service decision. This situation requires the following version of (R'):

(MR') Find $\underline{x}_k$ and $\underline{z}_k$, $k=1,2,\ldots,K$, in order to

$$\textit{Maximize } Z = \sum_{k=1}^{K} \underline{c}_k \underline{z}_k + \sum_{k=1}^{K} \underline{c}_k \underline{x}_k \qquad (3.7)$$

$$\textit{Subject to} \quad \sum_{k=1}^{K} A_k \underline{z}_k + \sum_{k=1}^{K} A_k \underline{x}_k \leq \underline{b} \qquad (3.8)$$

$$E_{jk} \underline{z}_{jk} + F_{jk} \underline{x}_{jk} \leq \underline{d}_{jk} \quad j=1,2,\ldots,J,\ k=1,2,\ldots,K. \qquad (3.9)$$

$$\underline{x}_{ijk}, \underline{z}_{ijk} \geq 0$$

where:

x_{ijk} is the level of an activity i in division j taking place <u>during</u> period k;

$\underline{x}_{jk}$ = $(x_{ijk}: i=1,2,\ldots,n_i)$, a vector of activity variables representing all n_i activities in division j during time period k;

$\underline{x}_k$ = $(\underline{x}_{jk}: j=1,2,\ldots,J)$, matrix of activity vectors representing all activities in all divisions during time period k;

z_{ijk} is the level of an activity i taking place <u>between</u> periods j and k;

$\underline{z}_{jk}$ = $(z_{ijk}: i=1,2,\ldots,m_i)$, a vector of all activities between time periods j and k;

$\underline{z}_k$ = ($\underline{z}_{jk}$: j=1,2,...,J), matrix of activity vectors representing all activities between period k and all other time periods;

$\underline{c}_{jk}$ = (c_{ijk}: i=1,2,...,n_i), a vector of contributions derived from all activities in division j during time period k;

$\underline{c}_k$ = (c_{jk}: j=1,2,...,J), matrix of all division j contribution vectors for time period k;

$\underline{b}$ = (b_1,b_2,...,b_K), a vector of resources available to the organization during each time period;

A_{ij} = (a_{ijk}:i=1,2,...,n_i), a vector of the amount of global resources used by division j during time period k;

A_k = (A_{ij}: j=1,2,...,J), a matrix of global resource useage vectors for all divisions during time period k;

E_{jk} = (e_{ijk}:i=1,2,...,n_i), a matrix of divisional resource usage coefficients for activities of division j during time period k;

F_{jk} = (f_{ijk}:i=1,2,...,n_i), a matrix of resource usage coefficients for all activities of all divisions between time periods j and k;

$\underline{d}_{jk}$ = a vector of resources available only to division j during each time period k;

X is the feasible region of $\underline{x}_{jk}$, i.e., the set to which allowable vectors $\underline{x}_{jk}$ are constrained.

Multiple time periods create separate subproblems for each divisional operation during each time period. Because the periods may be linked through variables that represent activities that occur between time periods, a new structure that includes both <u>coupling</u> <u>constraints</u> and <u>coupling</u> <u>variables</u> emerges. Coupling constraints are those that contain restrictions on the use of corporate resources by different divisions. Coupling variables relate entities across time periods. Where x_{ijk} might represent the decision to route an aircraft between patient service points i and j in time period k, z_{ijk} could be the decision to defer picking up a patient at point i in time period j until time period k.

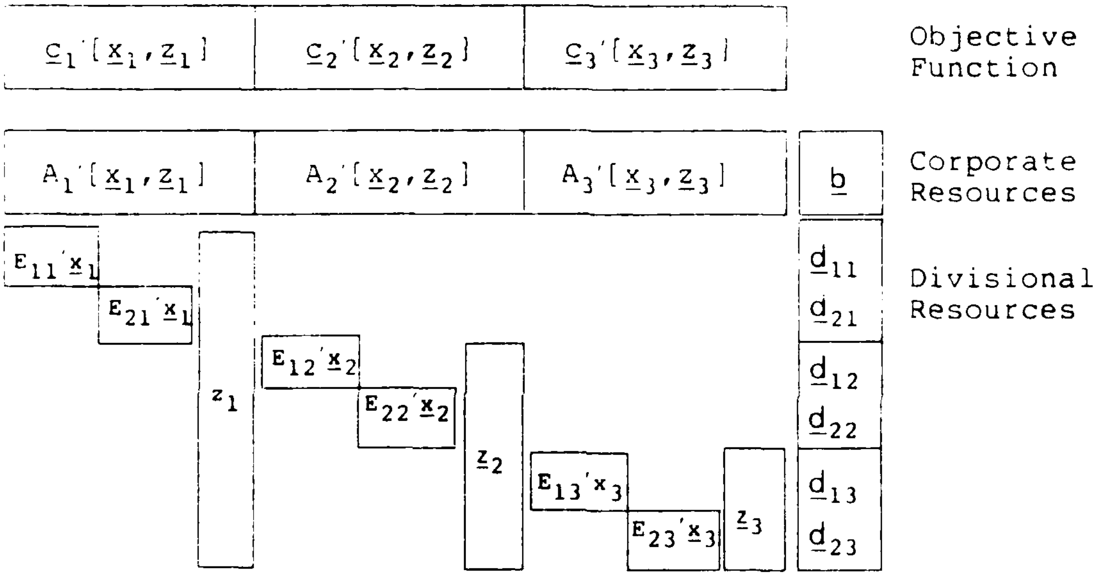


Figure 3.2. A three-period, two-divisional, decentralized structure.

Coupling variables create the structure shown in Figure 3.2. In addition to linking constraints containing limits on corporate resources, the rows of the divisional problems are now linked by variables that relate divisional operations across time periods. As Hillier and Lieberman [HILL81] have shown in multidivisional, multiperiod models, reordering the linking variables by moving the coupling variable columns to the left creates a *dual angular* structure in the divisional constraints, which is in turn the subproblem of a larger primal block angular problem with linking constraints.

3.4 Methods of Decomposing Large Mathematical Problems. Mathematically, the most interesting aspect of problems (R') and (MR') is that their structures permit a partitioning, or decomposition, into problems that can be solved without addressing the entire problem at one time. The theory of decomposition in mathematical programming is well developed, and a number of algorithms exist to solve problems that have the block angularity characteristic in their constraints. Algorithms generally use one of two types of coordinated information exchanges between decomposed subproblems, such that they yield the same optimal solution (if one exists) as would be achieved by solving the original problem directly. Although the dichotomy is admittedly weak,[3] the two principal coordination types are price and resource direction.

3.4.1 Price Direction. Dantzig and Wolfe [DANT61] reported the first price-directive technique,[4] based upon pioneering work by Koopmans, Kantorovich, Kuhn, Tucker and Hirschleifer. Baumol and Fabian [BAUM64] first interpreted the method in an organizational context. Essentially, pricing approaches achieve optimal allocations through the introduction of prices or penalities, issued by the superordinate unit (or "headquarters") for the use of global resources, in subordinate unit ("division") objective functions. (The terms Lagrange multipliers, dual variables and shadow prices are also used synonymously for prices.) In effect, prices impute the cost of the coupling constraints. The headquarters chooses pricing policies such that optimal feasible allocations eventually result. Divisions solve their individual problems in response to each successive pricing policy and transmit their requests for global resources to the headquarters for evaluation. Eventually, the headquarters obtains sufficient information to formulate the optimal pricing policy.

To demonstrate the functioning of the Dantzig-Wolfe model, consider problem (R') above. Define the set

$$S_k = \{\underline{x}_k \mid B_k\underline{x}_k \le \underline{d}_k,\ \underline{x}_k \ge 0\} \qquad (3.10)$$

as the feasible activities for division k, and assume S_k is bounded. Let $\underline{x}_k^e$ be an extreme point of S_k. Any $\underline{x}_k \in S_k$ can be written as a convex combination of extreme points:

$$\underline{x}_k = \sum_{e=1}^{E(k)} \lambda_k^e \underline{x}_k^e \qquad (3.11)$$

$$\lambda_k^e \ge 0 \qquad (3.12)$$

$$\sum_{e=1}^{E(k)} \lambda_k^e = 1 \qquad (3.13)$$

We can rewrite (R') as

(P')
$$\text{Maximize} \quad \sum_{k=1}^{K} \sum_{e=1}^{E(k)} (c_k' x_k^e) \lambda_k^e \qquad (3.4')$$

$$\text{Subject to:} \quad \sum_{k=1}^{K} \sum_{e=1}^{E(k)} (A_k x_k^e) \lambda_k^e \leq b \qquad (3.5')$$

$$\sum_{e=1}^{E(k)} \lambda_k^e = 1 \qquad (k = 1,2,\ldots,K) \qquad (3.6')$$

$$\text{all } \lambda_k^e \geq 0$$

(P') is called the restricted master, headquarters, or coordinator's problem. The decision variables are the $\underline{x}_k^e$. But, under the information structure assumption [JENN73] that X_k is not known outside division k, the extreme points $\underline{x}_k^e$ are not initially known to the headquarters. Dantzig-Wolfe uses supporting hyperplanes (column generation) to generate the extreme points. The headquarters issues $\underline{p}^t$, the shadow prices of (3.5'), where t is the iteration number. Division k then solves the following subproblem:

$$(P_k'(\underline{p}^t)) \qquad \text{Maximize } (\underline{c}_k' - \underline{p}^t A_k)\underline{x}_k \qquad (3.14)$$

$$\text{Subject to: } B_k \underline{x}_k \leq \underline{d}_k \qquad (3.15)$$

$$\underline{x}_k \geq 0$$

Let $\underline{x}_k^t$ be the optimal solution to $P_k'(\underline{p}^t))$. Division k transmits demand vector $A_k \underline{x}_k^t$ and payoff $\underline{c}_k' \underline{x}_k^t$ to the headquarters; transmitting $\underline{x}_k^t$ is unnecessary. [FREE73,p.65] If optimality is not reached, $\underline{p}^{t+1}$ is computed by solving (P'), and the process continues. It is well known that p_{opt}^t does not yield an extreme point solution for all divisions, and hence, the headquarters must impose the final solution on each division.

Detailed analyses of the various pricing approaches are given by Lasdon [LASD70], Geoffrion [GEOF70], Freeland [FREE73], Atkins [ATKI74], Ruefli [RUEF74], Molina [MOLI77], and Burton and Obel [BURT77]. Freeland derived the summary of representative algorithms in Table 3.3.

3.4.2 Resource Direction. In contrast to the price-directive approach, the resource-directive method, first proposed by Kornai and Liptak [KORN65,KORN67], partially reverses the roles of the headquarters and divisions.[5] (The first organizational interpretation is given in [BURT74].) Observing that the inability to decompose problem (R') into N independent subproblems is due to the global (or linkage) resource constraints, they proposed that headquarters iteratively allocate shares of resources to the divisions. With its current share (or tentative budget), each division solves its local problem and transmits shadow prices (also called bid prices or marginal values of increased budgets) on the use and value of the current tentative budget back to the headquarters. These prices are used to derive a new allocation, and the iterative process continues until a satisfactory near-optimum is reached.[6] ten Kate [TEN 72] developed a two-phased program (essentially the dual of the Dantzig-Wolfe method) that is guaranteed to achieve optimality in a finite number of steps.

TABLE 3.3

SUMMARY OF REPRESENTATIVE PRICING ALGORITHMS

Type of Model	Algorithm	Features	Complications
Linear	Dantzig-Wolfe (1958) [DANT61]	1. Primal feasible after a certain point. 2. Converges finitely.	1. Behavioral disadvantage: Headquarters must make final decision for subordinates.
	Balas (1966) [BALA66]	1. Prices are based on amount of infeasibility in primal problem. 2. Converges finitely. 3. Dual feasible (but not until convergence.	1. Behavioral disadvantage: Headquarters must make final decision for subordinates. 2. The headquarters must know the subordinates' bases matrices at each iteration.
	Jennergren (1972) [JENN72]	1. Uses a price schedule for each subordinate instead of a single price. 2. Converges finitely. 3. Primal feasible. 4. Subordinate's objective function quadratic.	1. The optimal dual multipliers must be known.
Quadratic (concave objective, linear constraints)	Haas (1968) [HAAS68] Whinston (1962) [WHIN62]	1. Handles simple externalities in the objective function. 2. Primal feasible. 3. If objective function is strictly concave with no externalities, optimal prices can be found.	1. Not necessarily finite. 2. With externalities the headquarters must make final decisions for subordinates.

Source: Freeland [FREE73]

Geoffrion [54] derived a commonly used taxonomy of problem manipulation and solution strategies for resource budgeting problems. The principal features of the three main approaches are given in Table 3.4. Computationally and behaviorally, there is no definitive way to establish the superiority of one method over another. We have chosen the tangential approximation approach for further discussion for a number of intuitively appealing reasons:

1. It requires the least amount of communication between levels.

2. Both the data required (resource shares and division solutions) and information flows between headquarters and divisions are what we might reasonably expect to either currently find in an organization or implement without major difficulty.

3. We expect it to perform particularly well when the feasible regions of divisional problems are easily described and objective functions well-defined.

The following section describes more rigorously the method of tangential approximation, summarizing the contributions of Geoffrion [GEOF70] and Freeland [FREE73]. The reader may only be interested in the general idea, which is given in the opening paragraph. The reader may then choose to go directly to the next section.

3.5 Resource-Directive Decomposition By Tangential Approximation. Essentially, we wish to decompose (R') into a series of subproblems corresponding to those of the individual divisions and the headquarters. (The same scheme

TABLE 3.4

THREE APPROACHES TO THE RESOURCE BUDGETING PROBLEM

Approach	Complications	Algorithms
1. Tangential Approximation: problem (R) is solved by iteratively building up an approximation for the objective function of (R). The tangential approximation is automatically available from the optimal multipliers of problem $R_k(\underline{b}_k)$ for a given $\underline{b}$.	1. Often hard to find feasible divisional allocations $\underline{b}_k$. 2. Most methods require the headquarters to know something about subordinate constraints.	Dantzig-Wolfe (1958) [DANT61] Kornai-Liptak (1967) [KORN67] Weitzman (1970) [WEIT70] ten Kate (1972) [TEN 72]
2. Large Step Subgradients: Because problem (R)'s objective is non-differentiable, large step gradient methods cannot be used. However the optimal multipliers of $R_k(\underline{b}_k)$ can be used to characterize $\underline{v}_k$ via subgradients, enabling optimal or near optimal or near optimal choices to improve $\underline{b}_k$.	1. Algorithms for the non-linear case require the headquarters to know something about subordinate constraints.	Abadie and Sakorovitch (1967) [GEOF70] Zschau (1967) [ZSCH67] Geoffrion (1970) [GEOF70] Silverman (1972) [SILV72]
3. Piecewise Approximation: Attempts to subdivide E^m into regions in which the $\underline{v}_k$ have a relatively simple structure and then solve problem (R) in piecewise fashion, each time with allocations restricted to one of these regions.	1. Unwieldy for problems that are not linear or quadratic. 2. Subordinates must transmit their current solutions and information about how they would change over a certain set.	Rosen (1964) [ROSE64] Varaiya (1966) [VARA72] Geoffrion (1970) [GEOF72]

Sources: Geoffrion [GEOF70], Freeland [FREE73] can be reiterated tɔ more than two levels when the divisions problems are themselves decomposable.) This is accomplished by creating a problem effectively equivalent to (R') using the mathematical techniques of projection (or partitioning), outer linearization, and relaxation. The headquarters, or master, problem generates tentative global resource allocations by approximating the response of each division to its tentative allocations. These approximations of optimal response are derived from the marginal (imputed) values on global and divisional resources reported by the divisions following subproblem optimization. The process terminates at optimality in a finite number of steps, or at least at a feasible, non-optimal solution. The creation of a master problem equivalent to (R') can be accomplished by projecting a two-variable system (with variables $\underline{x}_k$ and $\underline{b}_k$) onto the space of one variable ($\underline{b}_k$). To demonstrate this, suppose our problem is to

$$\underset{x\epsilon X;y\epsilon Y}{\text{Maximize}}\ f(x,y) \tag{3.16}$$

$$\text{Subject to: } G(x,y) \geq 0 \tag{3.17}$$

or graphically as in Figure 3.3. Projecting (3.14) onto the space of y alone, we have

$$\underset{y\epsilon Y}{\text{Maximize}}\ \{\underset{x\epsilon X}{\sup}\ f(x,y)\ |\ G(x,y)\geq 0\}. \tag{3.18}$$

Let v(y) equal the maximand in (3.18). The variable y must be in the effective domain V of y; otherwise v(y) is set equal to $-\infty$ to indicate infeasibility. Then

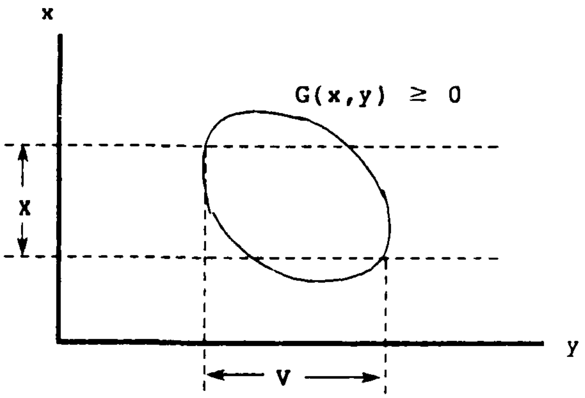


Figure 3.3. Depicting the set V.
Source:[GEOF70,p.116]

$$V = \{y : v(y) \geq -\infty\} \equiv \{y : G(x,y) \geq 0 \text{ for some } x \in X\}, \quad (3.19)$$

which makes our equivalent problem

$$\underset{y \in Y \cap V}{\text{Maximize}}\ v(y). \quad (3.20)$$

In our problem, we first introduce a modified constraint for (3.5) in (R'):

$$\sum_{k=1}^{K} A_k x_k \leq \underline{b}_k \quad (3.5a')$$

$$\sum_{k=1}^{K} \underline{b}_k \leq \underline{b} \quad (3.5b')$$

This revision changes (R') from a problem with coupling constraints to one with coupling variables. [GEOF70] We then project (R') onto the space of $\underline{b}$, which will

$$\text{Maximize} \sum_{k=1}^{K} v_k(\underline{b}_k) \quad (3.21)$$

$$\text{Subject to} \sum_{k=1}^{K} \underline{b}_k \leq \underline{b} \quad (3.22)$$

where

$\underline{b}_k \epsilon B_k = \{\underline{b}_k \epsilon E^m \mid B_k\ \underline{x}_k \le \underline{d}_k, A_k\underline{x}_k \le \underline{b}_k,$ for some $\underline{x}_k \epsilon X_k\}$, (3.23)

and $v_k(\underline{b}_k)$ is the optimal response $(\underline{x}_k{}^*)$ by division k to allocation $\underline{b}_k$. To find such responses we partition (R') into K subproblems that

$$(R''(\underline{b}_k)) \qquad \text{Maximize } \underline{c}_k{}'\underline{x}_k \tag{3.24}$$

$$\text{Subject to: } A_k\underline{x}_k \le \underline{b}_k \tag{3.25}$$

$$B_k\underline{x}_k = \underline{d}_k \tag{3.26}$$

$$\underline{x}_k \ge 0$$

A key point should be made here. The original problem (R') was transformed in to (R''). But, we cannot begin to solve (R'') directly because we do not know anything initially about $v_k(\underline{b}_k)$. This is the underlying reason for two modifications. First, before the initial iteration, the headquarters must be able to find a feasible allocation for all divisions. Past experience may provide it, or else some technique such as ten Kate's infeasibility form method [TEN 72] (based on the Dantzig-Wolfe Phase I algorithm) can be used. Secondly, through successive iterations, a series of approximations to $v_k(\underline{b}_k)$ must be developed. The method of tangential approximation uses outer linearization to accomplish this. (See Figure 3.4).

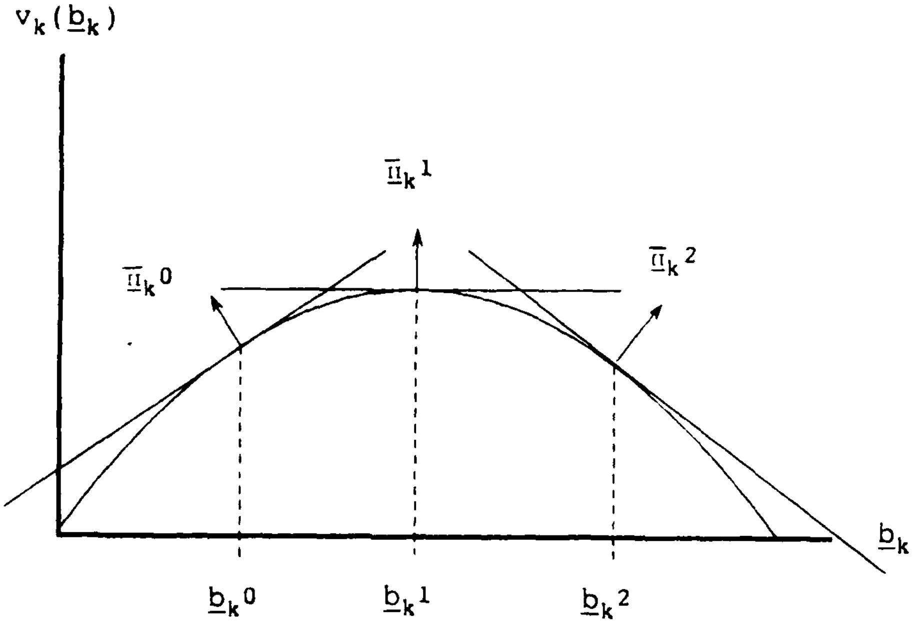


Figure 3.4. Approximating $v_k(\underline{b}_k)$.

Assume dual variables $\underline{\Pi}_k t$ exist for a given $\underline{b}_k t$. It can be shown that $\overline{\underline{\Pi}}_k t$ is a normal to $v_k(\underline{b}_k)$ at $\underline{b}_k t$ associated with the function

$$\underline{c}_k{}'\underline{x}_k - \overline{\underline{\Pi}}_k t(\underline{b}_k - \underline{b}_k t), \tag{3.27}$$

the tangent to $v_k(\underline{b}_k)$ at $\underline{b}_k t$. Therefore, each iteration provides an improved estimate to (R'') of the optimal response function v_k. ten Kate's [TEN 72] algorithm assures reaching optimality by requiring improvement in the approximation in each iteration.

In effect, then, the idea is for the divisions to provide the headquarters with imputations of resource values with which the headquarters can improve allocations and eventually achieve optimality. (Since the piecewise approximation never underestimates v_k, a monotonically decreasing upper bound on the optimal value of (R') is known. And, since each allocation is feasible, the best solution based on all previous allocations provides a lower bound, so that a measure of potential improvement is available.) Finally, the information requirements are minimized, since the headquarters only has to transmit allocation decisions, and divisions need only report resource shadow prices that are automatically generated when they solve their planning problems.

Decentralization has been shown in some models to be at best only partial. The price-directive method requires that the final solution be imposed upon the divisions, since the optimal solution is formed by weighting the divisional solutions. That is because the global optimal will not necessarily coincide with each divisional optimum. Jennergren's method [JENN72,JENN73] of using (linear) price schedules overcomes this problem. But, his divisional objective functions are quadratic, which increases computational complexity, and more importantly, only an infinite number of iterations assures convergence, which causes implementation problems. Other methods (e.g., the preemptive goal approach of Charnes, Clower and Kortanek [CHAR67]) resolve price inadequacy, but at the expense of increased complexity and added information transmission.

In the resource-directive case, Jennergren [JENN73] proves that feasibility throughout the solution process is not assured. To maintain feasibility in each iteration, each division must receive a resource allotment sufficient to assure existence of a feasible solution. However, for resource-directive methods, the property that causes infeasibility to occur (in some iterations) is the allowance of unbounded divisional problems, which is not realistic. Further, if successive planning periods are highly similar, initial starting strategies based on previous experience should ensure initial feasibility. However, no clearly superior initiation strategy under resource-direction has been reported. [BURT77,p.403]

A potentially serious computational problem arises in resource direction due to degeneracy. Freeland and Moore [FREE73,p.1053] define global resources required by every subordinate unit fully competitive. If, for decomposable linear programs that possess a feasible and a unique, bounded and non-degenerate optimal solution,[7] at least two divisions, and at least one global resource,[8] then:

1. All resource-directive decompositions are degenerate at optimality, i.e., partitioning global resources induces at least one degenerate optimal divisional solution.
2. At least g(n-1) basic variables are zero, where g and n are the numbers of global resources and divisions respectively.[9]

3. Only one divisional optimal solution may be non-degenerate, which implies that no more than one optimal bid vector is unique.

4. At optimality, the optimal bid vector is not, for all divisions, an extreme point of the set of all optimal vectors. [FREE77,p.1056]

Although ten Kate [FREE77,p.1053] proved that at least one common bid price vector for all divisions exists at optimality, degeneracy in a divisional optimal means that its set of vectors is not a singleton.[10] Therefore, there is some chance that optimality will not be achieved, particularly when all divisional problems are degenerate. The Kornai-Liptak model won't recognize optimality if the initial allocation is optimal and divisional problems are degenerate. If degeneracy elimination techniques are used (e.g., perturbation), equality of bids will not result. Assuming non-degeneracy (as Burton et al do [BURT74,p.303]) unrealistically avoids the issue. If the regularity condition that no alternative global optima exist is removed, then non-degeneracy may result. [FREE77,p.1054ff.]

The important implication is this. Using the well-known principle that "adjustments should be made until a resource yields equal marginal return in all uses," [BURT77,p.402] the headquarters' attempt "to find an optimal allocation is almost surely doomed to failure." [FREE77,p.18] More sophisticated algorithms can lessen the problems of degeneracy, but as Freeland and Moore [FREE77] point out, such algorithmic modifications add complexity and have no direct organizational analog.[11]

3.5.3 A Computational Example of Resource Direction. To demonstrate the resource-directive decomposition method, consider the following linear program:

$$\begin{array}{llllllllllll}
\text{Maximize } Z = & x_{11} & + & 2x_{12} & & & + & x_{21} & & & & \\
\text{Subject to:} & x_{11} & + & x_{12} & + & x_{13} & + & x_{21} & + & x_{22} & \leq & 10 \\
 & 4x_{11} & + & 2x_{12} & + & 3x_{13} & & & & & \leq & 10 \\
 & x_{11} & + & 3x_{12} & + & 2x_{13} & & & & & \leq & 10 \\
 & & & & & x_{13} & & & & & \leq & 5 \\
 & & & & & & & x_{21} & + & x_{22} & \leq & 15 \\
 & & & & & & & 2x_{21} & - & x_{22} & \leq & 20 \\
 & & & & & & & 2x_{21} & - & x_{22} & \leq & 10 \\
\end{array}$$

$$x_{ij} \geq 0, \quad i=1,2, \; j=1,3.$$

The block angular structure, with one coupling constraint for one global resource and two divisional problems, should be apparent. To obtain a resource direction solution, two types of linear programs must be formulated. One is the problem each division solves. For example, division 1 will

$$\begin{array}{lllllllll}
\text{Maximize } \varnothing_{1}^{k} = & x_{11} & + & 2x_{12} & & & & & \\
\text{Subject to:} & x_{11} & + & x_{12} & + & x_{13} & \leq & b_{1}^{1} & (U_{11}^{k}) \\
 & 4x_{11} & + & 2x_{12} & + & 3x_{13} & \leq & 10 & (V_{1}^{k}) \\
 & x_{11} & + & 3x_{12} & + & 2x_{13} & \leq & 10 & (V_{2}^{k}) \\
 & & & & & x_{13} & \leq & 5 & (V_{3}^{k}) \\
\end{array}$$

$$x_{ij} \geq 0, \quad i=1,2, \; j=1,3.$$

Division 2 solves an analagous problem. The U_{ij}^{k} and V_{i}^{k} are dual variables for the global and divisional resources during iteration k. The amount of global resource i allocated to division j is denoted b_{i}^{j}. The amount is determined by the master (or coordinating) problem. Each then reports the values of the dual variables, U_{ij}^k, for each global resource i allocated to division j during iteration k, and the the imputed value of its m divisional resources, $\Phi_j^k = d_{1j}V_1^k + \ldots + d_{mj}V_m^k$, during iteration k. This is all the information required by the master problem to coordinate the solution process.

To begin the solution process, we need an initial allocation of resources. Earlier, we said this can be done mathematically. We can also use our knowledge of solutions that yielded reasonable results in the past. Lacking any other alternative we can divide shares equally.

Suppose we allocate equal shares of 5 units to both divisions. Divisional solutions are $\varnothing_1^1 = 7$ and $\varnothing_1^1 = 5$, imputed divisional resource values of $\Phi_1^1 = 7$ and $\Phi_2^1 = 0$, and global resource shadow prices $U_{11}^1 = 0$ and $U_{12}^1 = 1$ respectively. Combined objective achievement is 12; division 2 did not fully utilize any divisional resource but used all of its global allocation, and just the reverse occurred in division 1.

When the headquarters receives the divisional shadow price information, it can then solve the following problem:

$$\begin{array}{lllll} \text{Maximize} & \theta_1 + \theta_2 & & & \\ \text{Subject to:} & \theta_1 & -U_{11}^k b_1^1 & & \le \Phi_1^k \\ & \theta_2 & & -U_{12}^k b_1^2 & \le \Phi_2^k \\ & & b_1^1 & + \; b_1^2 & \le 10 \\ & & k=1,2,\ldots,t-1. & & \end{array}$$

As the notation implies, the master problem adds one row for each division after each iteration. (For that reason a dual simplex solution method is often used.) The objective, $\sigma_1+\sigma_2$, is a monotonically decreasing upper bound on total objective achievement. The first solution revises the allocations to $b_1{}^1 = 0$ and $b_1{}^2 = 10$, primarily because the global shadow price reported by division 1 for the initial allocation was zero. With this new allocation, the two divisional problems are solved, yielding only one change: $U_{11}{}^2 = 2$. Two constraints are added to the master problem, the solution $(\sigma_1,\sigma_2,b_1{}^1,b_1{}^2) = (7,17/2,3/2,17/2)$ is found, and these new allocations are transmitted to the divisions. Eventually, the headquarters receives the same price vectors in succession, indicating the last allocation was optimal. The global optimal, $(x_{11}{}^*,x_{12}{}^*,x_{13}{}^*,x_{21}{}^*,x_{22}{}^*)$ $=(0,10/3,0,20/3,0)$, requires an optimal allocation of $(b_1{}^1,b_1{}^2)^* = (10/3,20/3)$. Progress toward optimality in the first iterations is evident in this example, as is a substantial fluctuation in allocations, although neither of these conditions is always present.

This example illustrates the principal value of the resource-directive method: resources are distributed in a series of trial allocations until one is found that puts them to best use. During that process, each division finds the best solution to its own problem that it can achieve with the allocation it receives for each iteration.

3.6 A Resource-Directive Decomposition Approach to DOD Domestic Aeromedical Planning. There are a number of properties of the resource-directive method that appear particularly germaine to the design and development of an aeromedical systems planning model. First, it allows a large problem to be broken down into smaller and more tractable parts, perhaps allowing even greater detail to be included in the subproblems. In principle, we could represent the interests of individual patients or small groups of them with, say, an origin or origin and destination in common. Secondly, we can directly observe the simultaneous, whole system effects of changing resource allocations on service to groups or individuals represented in the subproblems. And thirdly, these allocations can be made so as to achieve improvement on a coordinated basis, where the attentions of various corporate entities (headquarters, divisions, etc.) can be directed toward parts of the planning problem without jeopardizing the interests of the whole system.

This section addresses three design issues. First, we outline the salient features we ought to include in the model. Secondly, we discuss the literature that is relevant to this problem. As we will show, no model previously reported coincides exactly with the model we ought to design. And because of that, we will conclude with a plan of attack for developing the series of model versions discussed earlier that incorporate those features.

3.6.1 Aeromedical Planninq Model Specifications. Before we begin model construction, we first need to specify what we want it to accomplish. This includes not only characterizing the model components (objective functions, constraints, decision variables, etc.), but also the particular decision roles and functions we are trying to support. Our purpose in doing this, in addition to making implementation our foremost concern, is to establish a framework with which we can evaluate similar models in the literature, as we will do in the next section.

There are three major types of plans for which the aeromedical organization needs a model to provide support in making decisions. The first, which we call a daily routing plan, is to determine the best routes to visit a given set of patient origins and destinations and the staging facilities. By given we mean that through some means, the particular patients who will be transferred have been selected. We assume that the beginning and ending points of each route (or mission) are given by another plan, that we call the routing design. These end points are obviously important to the missions, but they should be chosen so that the best level of service is provided over as many time periods as possible. In analyzing patient flows, we observed that the flow patterns were quite similar between calendar weeks, so we have used a week as the planning horizon for routing design.

The third type of plan is for the movement of each individual patient. Those involved in developing a plan include the patient, the patient's physician, the originating medical facility administrator who authenticates the transfer, the ASMRO regulator who validates the transfer, and, at the destination hospital, the physician who accepts the referral, and the administrator who determines his facility's capacity and clinical capabilities to accept referrals. As we demonstrated earlier, one individual patient movement plan can significantly affect other patient movements. The capability to alter individual movement plans will not be included in the model, but should be addressed in future research. However, the final version of the model will indicate the impact of altering destination choices on the daily routing plan, information which could be used by regulation decision makers.

In developing these plans, the model should observe the restrictions discussed in the last chapter in achieving the best possible measure of performance in terms of client service. To formulate these constraints and objectives mathematically, individual resources (aircraft, crews, budget, etc.) can be combined to form a single resource unit in terms of which all of the most important restrictions we discussed in the section on system environment can be expressed. These units are route segments connecting staging facility bases and patient service points.

The general planning problem, then is to allocate these resources to achieve their best use in serving the system's clients. Based upon our discusion of performance measures, we presume that patients are best served if we

I. System Objectives:

A. Minimize pickup and enroute delays, and
B. Minimize enroute travel time.

In achieving these objectives, the model should

II. Constraints:

A. Patient service:

1. Not exceed the maximum time before pickup allowed by DOD rules;
2. Not exceed a stipulated maximum time in the system, from the time a patient is reported for movement until delivered to final destination;
3. Allow enroute overnight stops only at depots (aeromedical staging facilities);

B. Routing feasibility:

1. Begin and end aircraft routes only at the depots;
2. Observe pickup/delivery ordering;
3. Not assign more than a stipulated maximum number of segments to a route;

C. Operating restrictions

1. Observe maximum aircraft range and capacity;
2. Restrict maximum trip length (consecutive missions before returning to the central base).

Before we begin to mathematically formulate these constraints and restrictions, we will first review similar models reported in the literature to see how others have incorporated them in their models.

3.6.2 Literature Review. To meet our needs exactly, a model would have to address or incorporate the following aspects or concerns. First, this is a public sector problem, so the measures of performance would have to be different from private sector objectives, such as maximizing profit, and instead focus on the legitimate concerns of clients who benefit from a service provided by a public sector organization. Secondly, our principal focus is on medical transportation, so the considerable volume of work on moving commodities whose survival or comfort is not in question may not not relevant. The special features of the problem, the regional demand structure, the need to provide the service over a time horizon, and the need to preserve the identities and needs of each patient (at least in terms of discrete origins and destinations) must be included. And fourthly, a model should show explicitly the relationship between client service and the allocation of resource units.

In short, no model previously reported has the capability to handle the aeromedical problem exactly as we have described it. The literature on decomposition techniques, which is very extensive, is a case in point. Narrowing our focus to resource direction techniques used in public sector applications, a number of allocation mechanisms employing resource budgeting have been reported, as Table 3.5 shows.

TABLE 3.5

PUBLIC SECTOR APPLICATIONS OF RESOURCE BUDGETING

Author(s) (Year) [Source]	Application
Kornai and Liptak (1965) [KORN65]	Central planning in a socialist economy
Malinvaud (1967) [MALI67]	Central planning in a non-socialist economy
Wietzman (1970) [WIET70]	Central economic planning
Cassidy, Kirby and Raike (1971) [CASS71]	US Federal government revenue sharing with states and cities
Ruefli (1971) [REUF71]	US Department of Defense Programming-Planning-Budgeting System (PPBS)
Crecine (1970) [CREC70]	US Department of Defense Budgeting
Obel and Christensen (1976) [BURT77]	Regional model of Danish agriculture

None of these involve transportation. In fact, transportation applications of decomposition techniques are not common. In reviewing the literature for applications similar to the aeromedical problem, we could not find a single reference to a transportation model with geographical structure, multiple time periods, and multiple commodity movement, in a resource-directive decomposition framework. We did find three applications that were sufficiently similar to the aeromedical problem to warrant further investigation. The first features vehicle routing over time, the second a geographic regional organization requiring coordinated solutions, and the third multiple commodities moved in an air transportation network.

Agin and Cullen [AGIN75] developed a model for the purpose of modeling large-scale military deployments involving multiple commodities, multiple modes (multiple vehicles of different types) over multiple time periods. The purpose of their TRAVEL model was to determine, for each planning period, what routes each vehicle would take and what commodities would be moved by each of the routed vehicles, in order to minimize the cost of holding commodities at non-demand points and shortages at demand points. The network of allowable vehicle routes, route capacities, commodity availability and latest delivery times; and the number, location, and capacities of all vehicles are all assumed given. Commodities can be transshiped, and vehicles can carry different commodities simultaneously.

The TRAVEL model is solved by using a heuristic variant of the Dantzig-Wolfe price-directive technique called reflection programming, in which only one subproblem appears in the extremal problem at a time. The routings and loadings of each vehicle are iteratively improved until no further improvements can be found. Optimal solutions are possible, but not assured. The reflection programming technique has never been formally reported, so we could not evaluate it further. We do incorporate a number of constraints in our model similar to those in the TRAVEL model.

Among early efforts, the decentralized transshipment model formulated and solved by Ruefli [RUEF71] is perhaps most notable for its explicit treatment of organizational structure. Although as Davis [DAVI75] observes, Reufli's model can be viewed as a relatively simple extension of Dantzig's transshipment model [DANT63], Ruefli's purpose is to examine the possibility of representing organizational structures, such as the regions of a geographically large transportation network, that are controlled by separate decision makers working for a central manager. Davis extended Ruefli's model to the three levels shown in Figure 3.5.

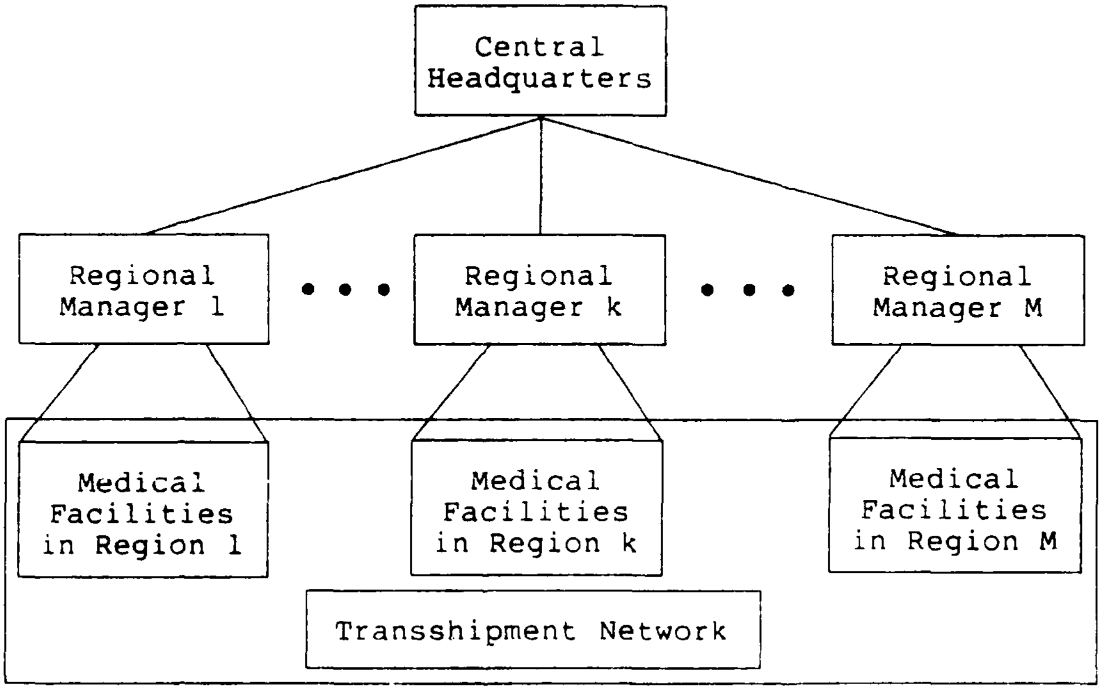


Figure 3.5. Ruefli's transshipment model.
Source: [DAVI75]

In Ruefli's model, the demand at each node is given, but both the supply at each node and the amounts transferred between geographic regions are decision variables in the central headquarters problem. His model is very flexible in terms of the costs or penalties that can be attached to interregional commodity exchanges. Once the amounts to be exchanged between regions (which are referred to as <u>imports</u> and <u>exports</u>) and the supplies are decided, the regional subproblems are solved for the minimum cost distribution plans. By reporting shadow price information to the central headquarters, the regions receive revised allocations of supplies and new import/export levels that improve whole system improvement (though individual regional achievement may decrease in some cases), until this coordination process reaches optimality. The process requires no communication between regions, and the solution of regional problems can be accomplished by the regional managers without central headquarters intervention. This is the concept of decentralization that Ruefli intends when he refers to this as decentralized transshipment.

Ruefli utilizes a single-commodity flow network with fixed node-arc structure. The actual means of flow distribution, i.e., the vehicles and their routes, are not explicitly modeled. To our knowledge, no one has reported a decentralized regional transshipment model incorporating the problem of feasible vehicle routing.

Kennington [KENN78] provides a state-of-the-art survey of the results of using linear multiple commodity network flow (MCNF) algorithms in distribution, routing and vehicle scheduling applications. The essence of these problems is to move distinct commodities from one or more supply points to one or more demand points over a fixed network with transshipment points, arc capacities, and flow conservation restrictions. Time periods can be explicitly included by replicating nodes in each period. He discusses three major solution techniques, including resource direction.

The basic idea of the resource-directive approach to MCNF problems is to distribute the limited capacity of arcs in a (fixed) network among the various commodities. To accomplish this, the K-commodity problem is decomposed into K single-commodity problems for which there are efficient solution techniques. Feasibility is insured by restricting the sum of the capacities of an arc over all commodities to be less than or equal to the arc's capacity in the original problem. Allocations are revised in each iteration on the basis of pricing information provided by the K sub-problem solutions, so that the process monotonically improves global objective achievement until optimality is reached.

Ali and Kennington [ALI 81] have successfully applied the MCNF approach to an air transportation problem in which distinct cargos must be shipped between 60 bases in the US.

Given the demands (in pounds) for each origin-destination base pair and the distances (in miles) between them, they first solve a MCNF problem in which the flow not only satisfies all demands at minimal cost (in pound-miles), but also occurs over collections of arcs that form circuits. These circuits represent nominal aircraft routes, but because aircraft capacity and other operating restrictions may be violated, these nominal routes are used to construct a set of feasible routes. A specialized form of MCNF, a muiitcommodity fixed charge network model is used to select the optimal set of routes that that will guarantee that all demands and all aircraft operating restrictions are met.

Ali and Kennington employed a partitioning technique to solve the air cargo problem. Kennington is currently developing a resource-directive code for the MCNF problem, but he has not reported it formally. The resource-directive code is expected to overcome the size limitations experienced with the special primal basis partitioning code used in the air cargo model, which is nonetheless very efficient for small problems.

The MCNF applications by Ali and Kennington are all single period models. They do not incorporate explicit organizational structures. The air cargo model does employ multiple vehicles, and requires them (the aircraft) to follow circuits that return them to their "home stations".

Swoveland [SWOV71] used an MCNF formulation and a resource-directive solution technique to solve a multi-commodity, multi-period distribution problem. He assumed a fixed network to model a production-inventory-distribution problem over several time periods. His solution times were disappointing when compared with price-directive methods, but Kennington [KENN78] attributes this to the choice of techniques used to construct his resource-directive algorithm. Swoveland's model does include multiple time periods, which he models by creating replicates of nodes in each period, and he maintains the production, inventory and distribution functions as separate subnetworks within the overall network. His model does allow multiple commodities, but it does not consider vehicle routing as part of the distribution problem.

3.6.3 Approach. Since none of these models incorporates all of the objectives and constraints specified in the last section, it was apparent that a new model was needed. In the following chapters, we will follow the outline given in Table 3.2 to develop four versions of the model. We will first examine the most fundamental aeromedical transportation problem, routing a single aircraft through a set of patient service points. This first model assumes we must serve all patients and that we have no resource restrictions. The model will be formulated as a variant of a classic integer linear program called the traveling salesman problem (TSP), with additional constraints added to insure that routes visit patient destinations only after visiting their origins. The solution method uses a branch-and-bound technique and employs a number of recently developed devices to increase computational efficiency. We will then extend the single aircraft model to include multiple depots and multiple aircraft. The extension is relatively straightforward.

The complexity of the problem increases greatly when we attempt to include additional problem characteristics, such as allowing partial patient service in one time period while guaranteeing complete patient service over several time periods. We also require the aircraft to cycle through the central base periodically. Because of the complexity, we have to exploit the hierarchical structure of patient movements within and between regions. Finally, we will use a resource-directive approach to solve weekly schedule problems that allows us to break down the very large problem that result into a much smaller one. In addition to allowing a feasible solution to be found, the decomposition approach generates information about the changes in service among clients that result when resource allocations are changed, a potential we observed in our computational demonstration.

ENDNOTES

1. The idea is that divisional resources are owned, immobile or otherwise not transferable, or that a separate allocation decision has already occurred. In the case to be studied, for xample, resources such as the staging facilities are assumed fixed, because transfers involve major political and logistical difficulties as well as substantial resource expenditures. Theoretically, such transfers could be easily handled in (R') by relabeling transferable resources as global.

2. Ruefli [REUF71b] and Watne [WATN77] define two types of externalities treated in the decomposition literature. Behavioral externalities are defined as "the interdependencies that arise when there is a behavioral (e.g., psychological) relation between the efforts of one management unit to reach its goal levels and the efforts of a different management unit to reach its own goal levels." [136,p.30] Three types of technological externalities have been dealt with by several authors. Watne's thesis [WATN77] deals extensively with the situation where the activity level of a management unit is dependent on the level of another kind of activity in another management unit. Technological externalities create constraints of the type $g_k(\underline{x}_k,\underline{x}_l)\geq 0$, which requires the introduction of new variables and constraints z and $\underline{x}_l-z=0$. This has the effect of internalizing the externality. A multiple

pricing problem arises if $\underline{x}_1$ affects more than one management unit. [REUF74,p.355]

3. As many authors have noted, some models, for example, Ruefli's GGD model [RUEF69], use both types.

4. Freeland [50] calls price direction *coordination through goal intervention*. Other common synonyms include *transfer pricing* [ABDE74] and *indirect distribution* [TEN 72]. Hirschleifer's article [HIRS57] is the earliest known treatment of the subject.

5. Resource direction is also called *coordination through constraint intervention* [FREE73], *direct distribution* [TEN 72], and *resource budgeting*.

6. The Kornai-Liptak algorithm does not converge in a finite number of steps. The problem stems from the quasi-linearity of the divisional objective functions: within regions of the resource-share space the objective is linear, but if shares are changed beyond limited values unknown to the headquarters, shadow prices change instantaneously. Kornai and Liptak utilize arbitrary bounds and an averaging technique, rather than mathematically rigorous criterion, to revise allocations such that convergence is not assured. [TEN 72,p.885]

7. Non-degenerate solutions are those in which all basis (solution) variables are non-zero. Such solutions yield a single optimal bid price vector.

8. Freeland and More assume all global variables are fully competitive only for ease of explanation. They call such L.P.'s _regular_.

9. This is termed the _order-of-degeneracy_.

10. Freeland and Moore contend that the set is infinite for each division with degeneracy; the reference they cite (Eilon and Flavell [EILO74]) does not prove this assertion. It is true that many-sided shadow prices occur with degeneracy, and the point of all divisions choosing the same vector being probabilistic is still valid. We assume the infinite set can be created by taking convex combinations of the finite set.

11. ten Kate avoids difficulty by forcing improvement in the solution each iteration. However, the cost of doing so is increased information transmission (divisional resource shadow prices must also be sent) and increased complexity (in the master program, where a constraint must be added for each bid vector received.) [TEN 72,p.896-897]

CHAPTER IV

THE SINGLE VEHICLE, MANY-TO-MANY ROUTING PROBLEM

4.1 Introduction. This chapter describes an algorithm that solves one of the most basic aeromedical decision problems, routing a single aircraft from a staging facility base (depot), through all patient origins and destinations in a region, and back to the depot. We will define and examine methods to solve the vehicle routing problem (VRP), the general class to which this problem belongs, then mathematically formulate the ordering restriction to deliver patients after picking them up. After examining solution approaches in the literature, we will present a new algorithm for the single vehicle, pickup and delivery routing problem, and conclude with a computational demonstration. While other algorithms solve this problem, none solves both the single vehicle problem and its multiple vehicle extension (which we cover in the next chapter) exactly.

4.2 The Fundamental Problem of Vehicle Routing. Using the following definitions,

> A vehicle route is a sequence of pickup and/or delivery points which the vehicle must traverse in order, starting and ending at a depot or domicile. A vehicle schedule is a sequence of pickup and/or delivery points together with an associated set of arrival and departure times. The vehicle must traverse the points in the designated order at the specified times. [BODI8i, p.97]

Bodin and Golden [BODI81] define two fundamental problem classes:

> When arrival times at nodes and/or arcs are fixed in advance, we refer to the problem as a scheduling problem. When arrival times are unspecified, the problem is a straightforward routing problem. [BODI81, p.98]

The two classes are not necessarily mutually exclusive; two complicating characteristics, precedence relationships and time windows, can create combined routing and scheduling problems. [BODI81] Precedence relationships stipulate that one entity must be serviced before another, such as when an entity must be picked up before it is delivered, or when one entity has priority over another. Time windows prescribe when, relative to some time base, an entity must be serviced. Different types of windows include: a time interval $[T_1,T_2]$, where T_1 and T_2 are acceptable service commencement times; a fixed service commencement time $(T_1=T_2)$; a one-sided interval $(T_1=-\infty$ or $T_2=\infty)$; or windows based upon service completion, rather than commencement.

The aeromedical problem includes precedence relationships, but neither time windows nor fixed service times. We are primarily interested, then, in finding a sequence of visits to all patient service points, subject to constraints, that maximizes or minimizes some objective measure of performance. The nature of the medical service provided to the patient is usually not germane to the routing problem, but a patient's medical needs may impose constraints such as no intermediate stops.

A VRP, depending upon the specific application, may incorporate one[1] of a number of objectives, such as:

1) minimizing routing (time, distance or monetary) costs;

2) minimizing routing and vehicle acquisition costs;

3) minimizing the number of vehicles required;

4) minimizing customer disutility (e.g., a function of waiting and riding time) when the service involves transporting the customer.

Our discussion of vehicle routing problems will make use of the graph-theoretic concepts introduced in Chapter 2. Given the set of nodes $N = \{1,2,...,n\}$, $n = |N|$, indexing n-1 cities and a central depot, let the set of arcs between nodes be $A = \{1,2,...,a\}$, $a = |A|$, and $C = [c_{ij}]$ be the costs of traveling arcs (i,j). In vehicle routing problems, the solutions, $X = [x_{ij}]$, consist of a set of one or more paths, or routes, that vehicles will travel; the (binary) decision variable, x_{ij}, will equal one if arc (i,j) is in the solution, and zero otherwise.

4.3 The Single Vehicle Routing Problem (SVRP). The simplest SVRP (of which Version I of the aeromed model is an example) is the classic Traveling Salesman Problem (TSP), which is to find a route that starts at a central depot, visits n cities, and returns to the depot, and minimizes total distance traveled. To solve the problem, which has neither routing limits nor vehicle capacity constraints, we must find the matrix $X=[x_{ij}]$ that will:

Problem TSP:

$$Minimize\ Z = \sum_{i=1}^{n} \sum_{j=1}^{n} c_{ij} x_{ij} \tag{4.1}$$

$$subject\ to: \sum_{i=1}^{n} x_{ij} = 1, \qquad j = 1,2,\ldots,n; \tag{4.2}$$

$$\sum_{j=1}^{n} x_{ij} = 1, \qquad i = 1,2,\ldots,n; \tag{4.3}$$

$$X = [x_{ij}] \in T \tag{4.4}$$

$$X = 0\ or\ 1, \qquad \begin{array}{l} i = 1,2,\ldots,n, \\ j = 1,2,\ldots,n. \end{array} \tag{4.5}$$

Problem TSP can be interpreted as follows. Because the variables x_{ij} must be either zero or one, the objective Z sums the lengths of all arcs traveled. Constraint (4.2) requires that one and only one arc terminate at each node. Similarly, constraint (4.3) ensures that exactly one arc departs each node. The objective (4.1), together with constraints (4.2), (4.3), and (4.5), are the familiar assignment problem (AP). If we prohibit self-loops, then $c_{ii} = \infty$ for all i. With this restriction, AP becomes the Modified Assignment Problem (MAP).

In problem TSP, solutions must be tours, which are connected elementary spanning circuits of a complete graph G. More specifically, we seek minimum weight tours, where A is in the domain of d, the range of d is ≥ 0, and the weight of a tour is the sum of the arc weights, c_{ij}, of the arcs included in the tour. The set T in (4.4) prohibits solutions with subtours (tours connecting fewer than n nodes). Alternative ways to express T include: [GOLD77]

$$T = \{\{x_{ij}\}: \sum_{i \in Q} \sum_{j \notin Q} x_{ij} \geq 1 \ \forall \ \emptyset \neq Q \subset N\} \qquad (4.4a)$$

$$T = \{\{x_{ij}\}: \sum_{i \in Q} \sum_{j \in Q} x_{ij} \leq |Q| - 1 \ \forall \ Q \subseteq \{2,3,\ldots,n\}\} \qquad (4.4b)$$

$$T = \{\{x_{ij}\}: y_i - y_j + n \leq n-1 \ \textit{for} \ 2 \leq i \neq j \leq n \ \textit{for some real numbers} \ y_i, y_j\} \qquad (4.4a)$$

Note that |Q| is the number of elements (cardinality) of the set Q. The set $\bar{Q}$ is the complement of Q, i.e., $Q \cap \bar{Q} = \emptyset$.

To interpret each of the alternatives for expressing T, let |N| = 5, and choose Q = {1,2}. Suppose we have the solution $x_{12} = x_{21} = x_{34} = x_{45} = x_{53} = 1$, shown in Figure 4.1. This solution could not belong to T in (4.4a) because no arc (1,j) or (2,j), where $j \in \{3,4,5\} = \bar{Q}$, connects a node in Q with a node in its complement. 2^n constraints of type (4.4a) are required to prohibit all possible subtours.

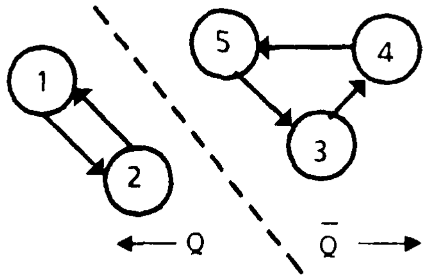
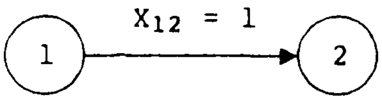


Figure 4.1. Solution to a TSP.

Constraints (4.4b) require that subsets of r nodes must have strictly fewer than r arcs connecting them. If we define Q as above, then the partial solution shown in

Figure 4.2 satisfies (4.4b), while the one in Figure 4.3



Fiqure 4.2. Feasible connection of two nodes.

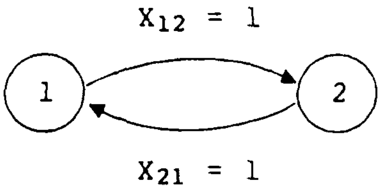


Figure 4.3. Infeasible interconnection of two nodes.

does not. Again, 2^n constraints are required.

Miller, Tucker and Zemlin [MILL60] devised the constraints in (4.4c). Let 1 index the depot. Select n^2 strictly positive, arbitrary real numbers u_i and u_j. Let the subtour $S_t=((i_2,i_3),(i_3,i_4),...,(i_p,i_2))$, where $p < n-1$ and $x_{i_2 i_3} = x_{i_3 i_4} = ... = x_{i_p i_2} = 1$. Then we can write the p - 1 constraints:

$$u_{i_2} - u_{i_3} + n \le n-1$$

$$u_{i_2} - u_{i_3} + n \le n-1$$

$$\vdots$$

$$u_{i_p} - u_{i_2} + n \le n-1$$

If we take a linear combination of these constraints,

$$\sum_{j=2}^{p} (u_{i_j} - u_{i_{j+1}} + n) \le (p-1)(n-1),$$

then $(p-1)n \le (p-1)(n-1)$, a contradiction. However, for a tour, $t = ((i_1,i_2),(i_2,i_3),\ldots,(i_n,i_1))$, choose $u_i = j$, $(j=1,2,\ldots,n)$, where node i is the ith node visited. If $x_{ij} = 1$, then $u_i - u_j + n < n-1$. Since $u_j = u_i+1$, $u_i-u_j = u_i-(u_i+1) = -1$; therefore, $n-1 \le n-1$. If $x_{ij} = 0$, $u_i-u_j \le n-1$, because $u_i < n-1$ for $u_j \ge 2$.

Any one of these three alternative constraints is sufficient to ensure that only tours can be feasible solutions to Problem TSP. Most solution methods, however, do not directly invoke these constraints as part of the solution process, primarily because of the large number of constraints involved. Rather, through a process known generally as relaxation, subtour prevention constraints are ignored (relaxed), the remaining problem (MAP) is solved, and its solution is checked for the presence of subtours. Using a recursive method known as divide-and-conquer, or more commonly as branch-and-bound, MAP solutions are modified to partially correct violations of constraint (4.5), and the relaxation process repeated, until the optimal solution is found. The next section will examine branch-and-bound TSP solution methods.

4.3.1 Branch-and-Bound TSP Solution Methods and Techniques.

In general, like dynamic programming, branch-and-bound methods are strategies, not algorithms; they must be altered to conform to the structures of specific problems. As Thesen notes,

> The strategy is based on the premise that the problem to be solved has the following attributes:
>
> (1) Combinatorial nature--a combinatorial problem has at the minimum the following properties:
>
> (a) A finite set of objects is given.
> (b) Each object can take on a certain range of attributes.
> (c) A solution to a problem is developed by fixing the attribute values for all objects.
> (d) Only certain combinations of attribute values are allowed.
>
> (2) Branchability--implies that
>
> (a) it must be possible to construct a finite and countable set containing all the different solutions to the problem (this follows from 1);
> (b) it must be possible to recursively partition a nonempty set of solutions into a nonoverlapping subset.
>
> (3) Rationality--a problem that
>
> (a) has solutions that yield a unique value calculated from the values of its attributes;
> (b) has a "best" solution that has the highest (or lowest) value.
>
> (4) Boundability--an estimate of the value of the best solution contained in any set of solutions can be obtained such that:
>
> (a) the actual value of the best solution in the set is inferior or equal to the estimate (thus, the estimate is a bound);
> (b) minimal effort is expended in obtaining this estimate;
> (c) the estimate is reasonably close to the actual value. [THES78,pp.171-172]

In Problem TSP, the objects are the decision variables, x_{ij}, with one (binary) attribute. Like many combinatorial problems, Problem TSP is NP-complete, with no known efficient solution methods. For that reason, the boundability attribute of Problem TSP is particularly important, since efficient MAP methods are available to generate estimates (bounds). Indeed, as an integer linear program, Problem TSP cannot be solved for sufficiently large n, and an indirect method such as branch and bound, or a heuristic, must be used.

For a formal treatment of branch-and-bound methodology, we strongly recommend Mitten's article. [MITT70] To understand our treatment of branch and bound methods, the reader should be familiar with the following concepts, and with assignment problem solution methods. Branch-and-bound methods construct a solution tree describing all MAP solutions explicitly solved, and implicitly enumerate large groups of solutions without explicitly solving them, through a process called fathoming or pruning. The nodes of the tree represent versions of the original problem with modifications to the data (costs) of the problem. The original MAP is the first, or root, node of the tree. We refer to the directed arc connecting two nodes as a branching, the initial node of a branching as the predecessor of the terminal node, and the terminal node as the successor of the initial node.

A branching represents a modification of the predecessor subproblem. In the remainder of the thesis, modifications are restricted to requiring a decision variable x_{ij} to have a value of 1 or 0 in the successor subproblem. (This corresponds to requiring or prohibiting travel between two cities). Therefore, if a branching requires one or more arcs to be traveled, the x_{ij} values corresponding to these included arcs must equal 1 in the subproblem solution, while prohibiting travel on an arc requires that the value of x_{ij} corresponding to the excluded arc must equal 0. Arc inclusions and exclusions along the path of branchings from the root node to a particular subproblem are cumulative; that is, a given subproblem retains all inclusions and exclusions of all of its predecessor problems.

Feasible solutions that satisfy (4.5) are tours. The upper bound is the lowest MAP objective value of any tour solution in the tree. The incumbent is a solution corresponding to the upper bound. Fathomed subproblems are those for which the solution (i) is a tour, or (ii), has an objective value greater than the upper bound. No descendents of fathomed subproblems are created; further branchings would create restrictions that could not possibly permit objective improvement. Unfathomed subproblems, those with infeasible solutions and objectives less than the upper bound, must be further branched. The lower bound is the lowest objective value of any unbranched subproblem. (Any subproblem objective is a lower bound to all of its descendents).

Most branch-and-bound TSP solution methods use the assignment relaxation technique to solve subproblems. The following operations for solving a subproblem are typical. All branching modifications are made to the cost matrix. The Hungarian or some other algorithm solves the assignment problem, and produces (i) the solution values of the x_{ij}'s, (ii) the objective value of the solution, and (iii), the final reduced matrix. If the solution is a tour, the upper bound and incumbent are changed (if the objective value is less than the upper bound), and the subproblem is fathomed. Otherwise, the (infeasible) objective value is compared with the upper bound. If less, the subproblem is unfathomed; if greater, it is fathomed.

The process selects and branches unfathomed subproblems until no unfathomed problems remain. Branching strategies prescribe subproblem selection. A best bound strategy selects the unfathomed subproblem with the lowest objective. A breadth strategy solves all the descendents of a predecessor. A strategy known by a variety of names such as depth-first, newest bound, or LIFO, solves the first descendent of the last problem solved. There is no conclusive evidence of the superiority of any strategy.

Regardless of the particular strategy employed, the object of branching is to construct mutually exclusive subsets of all solutions. If, among all possible branching paths, each solution can only be found on one branch, then the mutually exclusive and collectively exhaustive collection of subsets is called a partitioning. Branching strategies observe the well-known fact that it is necessary for optimal solution variables to have coefficients of zero in the final reduced cost matrix, but that it is not sufficient for a variable to be in the optimal basis if its final reduced coefficient is zero. Therefore, branching modifications are implemented as follows. Arc exclusions are forced by assigning an extremely large value (M) to c_{ij}, such that c_{ij} can never be reduced to zero. Assigning M to all row i and column j elements except c_{ij} will ensure that c_{ij} is reduced to zero and that arc (i,j) is included.

Researchers have proposed two basic types of branch and bound solutions for Problem TSP, tour building and subtour elimination. We will next examine both approaches,[2] and then present an extension to the latter that will solve the mixed service problem. Because upper and lower bounds play a particularly important role in determining the solution time of branch and bound algorithms, we also discuss some particularly effective bounding techniques.

4.3.1.1 The Tour Building Approach. The Little [LITT63] algorithm constructs tours by deciding in each iterative step if an arc is to be included in or excluded from the latest solution. The algorithm uses a best bound strategy to select the unfathomed subproblem to branch, chooses a branching variable, and creates two subproblems. In one, the arc associated with the branching variable is included in the solution, and in the second, the arc is excluded. Thus, each binary branching partitions the solution space, since arc inclusion and exclusion are mutually exclusive.

To construct the binary partition, the method utilizes an opportunity cost strategy to select the branching variable. Candidates for the branching variable are all variables with zero reduced cost in the final reduced cost matrix $[c_{ij}]$ of the branched subproblem. If candidate x_{ij} is not selected, then any solution without arc (i,j) will have to contain two arcs, one from i to any node other than j, and another from any node other than i to j. To make this selection, they first compute

$$e_{ij} = \min_{k:k \neq j} \{c_{ik}\} + \min_{k:k \neq i} \{c_{kj}\} \qquad (4.6)$$

for all variables when $c_{ij} = 0$. They then select the best branching variable, $x_{ij}^{*} = \max_{(i,j) \in A} \{e_{ij}\}$. Little et al show that this creates the largest increase in the lower bound from the lower bound of the parent problem. [LITT63]

Once x_{ij}^* is chosen, create two subproblems in which lb_y is the lower bound estimate of subproblem y, and h_z is the sum of reducing constants necessary to have at least one reduced cost of zero in every row and column of the final reduced cost matrix of subproblem z.[3] Fathoming tests are applied, and the process repeats with the lowest subproblem. To illustrate, suppose we have the cost matrix in Figure 4.5. If each matrix row is reduced by the

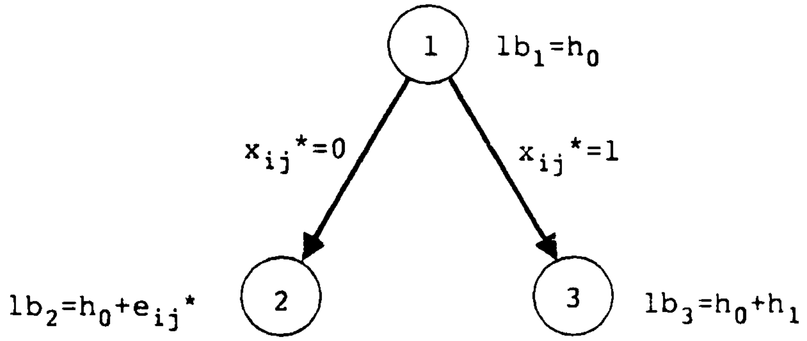


Figure 4.4. Little et al opportunity cost branching.

	1	2	3	4	5	6	7	8	9	10
1	M	184	292	449	670	516	598	618	881	909
2	184	M	195	310	540	357	514	434	697	964
3	292	195	M	215	380	232	434	493	719	955
4	449	310	215	M	288	200	566	787	790	1020
5	670	540	380	288	M	211	436	814	632	974
6	516	357	232	200	211	M	381	642	697	952
7	598	514	434	566	436	381	M	295	224	541
8	618	434	493	787	814	642	295	M	320	341
9	881	697	719	790	632	697	224	320	M	318
10	909	964	955	1020	974	952	541	341	318	M

Figure 4.5. Gillette's 10-city TSP cost matrix.
Source: [GILL76a]

smallest constant in it, and the resulting matrix columns 3, 5, 8 and 10 are further reduced by the constants 11, 11, 23 and 46 respectively, the final reduced matrix for the root subproblem results (Figure 4.6).

The superscripts in Figure 4.6 are the e_{ij}. Since $e_{12}^* = 97$, x_{12}^* is the branching variable. The first three subproblems, then, would be those shown in Figure 4.7. Note that for the arc exclusion branching, $lb_2 = lb_1 + e_{12}^* = 2346 + 97$. lb_3 is determined by deleting row 1 and column 2 from the final reduced cost matrix of subproblem 1 (Figure 4.6), setting $c_{21} = M$,[4] and summing the constants needed to reduce the rows and columns of the resulting matrix. In this example, row 2 and column 1 of that matrix must be reduced by 20 and 77 respectively. Since both subproblems 2 and 3 are unfathomed with the same lower bounds, one of them must be selected arbitrarily and the reduction process repeated.

There are several noteworthy features of the Little method. First, an upper bound can be found by making n successive arc inclusion branchings. Second, once an arc has been included, the associated row and column of the coefficient matrix can be deleted, because inclusion fixes the value of x_{ij} at 1. This reduces the size of the assignment problem that must be solved. Third, the method can handle problems with additional constraints, by restricting the candidates for branching variable to those

	1	2	3	4	5	6	7	8	9	10
1	M	0^{97}	97	265	475	332	414	411	697	779
2	0^{97}	M	0^{4}	126	345	173	330	227	513	734
3	97	0^{20}	M	20	174	37	239	275	524	714
4	249	110	4	M	77	0^{4}	366	564	590	776
5	459	329	158	77	M	0^{77}	225	580	421	716
6	316	157	21	0^{20}	0^{77}	M	181	419	497	706
7	374	290	199	342	201	157	M	48	0^{48}	271
8	323	139	187	492	508	347	0^{0}	M	25	0^{48}
9	657	473	484	566	397	473	0^{83}	83	M	48
10	591	646	626	702	645	634	223	0^{48}	0^{0}	M

Figure 4.6. Final reduced cost matrix for the root subproblem.

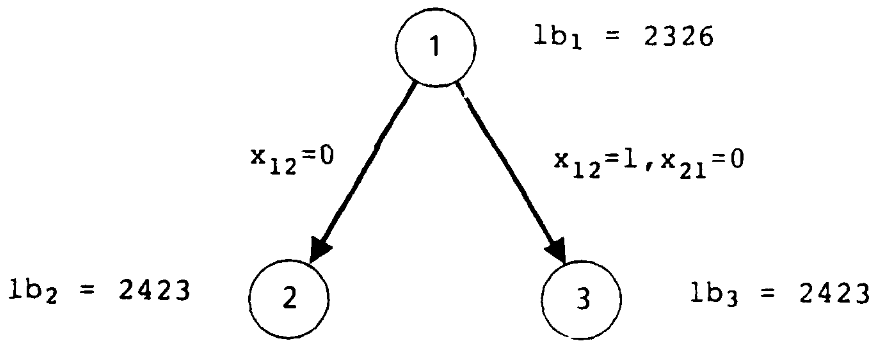


Figure 4.7. Partial solution tree using Little's algorithm.

which would not violate the additional constraints. In other methods, treating violations of additional constraints requires more extensive algorithm modification.

4.3.1.2 The Subtour Elimination Approach. Subtour elimination methods implicitly enumerate solution space subsets in a manner similar to tour building procedures. However, the subtour elimination approach differs in the way subproblems are branched. Where tour building ignores the solution if it is not optimal, and uses an opportunity cost analysis of the final reduced matrix to select the branching variable, subtour elimination attempts to induce feasibility by partially prohibiting the conditions that caused the infeasibility (subtour) in a subproblem assignment solution. Several different methods have been proposed for doing so.

Bellmore and Malone propose the following general subtour elimination method:

Step 1. Solve the MAP.

Step 2. Check for subtours.

Step 3. Eliminate subtours by imposing conditions on the solution subspace that do not eliminate feasible solutions.

Step 4. Repeat 2 and 3 until an optimal feasible tour is found.[BELL74]

If we use the number of articles in the literature as the criterion, it appears that subtour elimination methods are the most widely implemented solution method. In the following sections, we will discuss subtour elimination procedures in the chronological order of their development.

4.3.1.2.1 Eastman-Shapiro. The first reported branch-and-bound subtour elimination method was originally proposed by Eastman [EAST58], and later modified by Shapiro. [SHAP66] The central feature of the method is the elimination of one subtour in an infeasible solution by excluding arcs in that subtour. To illustrate, suppose we have found the solution to a six-city TSP depicted in Figure 4.8.

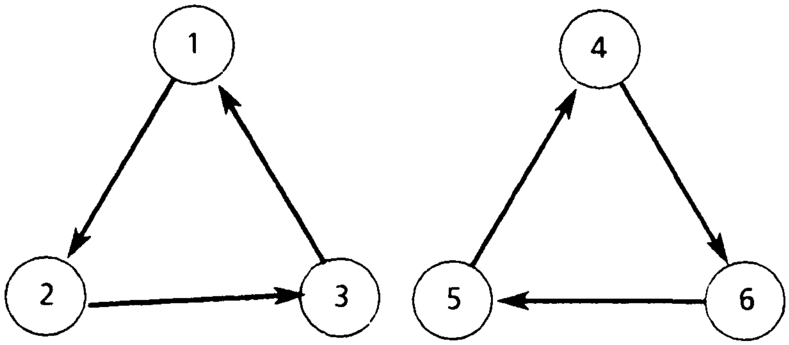


Figure 4.8. Six-city TSP solution.

Let S = {1,2,3} be the set of nodes in the shortest[5] subtour. Let S′ = {(1,2),(1,3),(2,1),(2,3),(3,1),(3,2)} be the set of arcs that could make {S,S′} a connected graph. That is, let S′= {(i,j)|i,j∈S}, and k = |S|. Bellmore and Malone have proven that, "Imposition of the constraint, $\sum_{(i,j)\in S'} x_{ij} \le k-1$, "either by integer programming or by splitting the solution space, eliminates all solutions that involve the subtour S". [BELL74] This theorem establishes constraint (4.4b). A subtour exists because the number of arcs and nodes in the circuit are equal. Eastman and Shapiro eliminate at least one subtour by prohibiting one arc of the shortest subtour in each descendent subproblem. That is, they create k subproblems, each prohibiting a different arc of the shortest subtour.

As Figure 4.9 shows for our example, three new MAP's are formed. No solutions are prohibited; e.g., subproblem

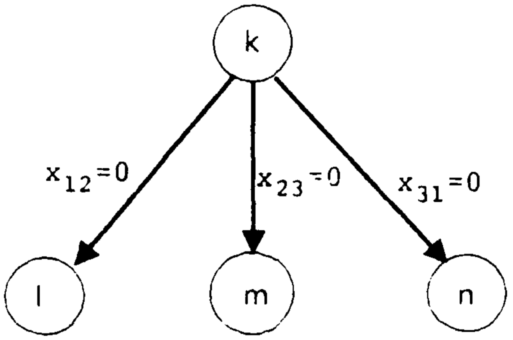


Figure 4.9. Eastman-Shapiro branching.

l and its descendents can not have arc (1,2) in any solution, but neither subproblem m nor n prohibit that arc. Arc exclusions only guarantee that one subtour will be eliminated from l, m, and n; the subtour can occur in other non-descendent subproblems in the solution tree, and solutions of descendents l, m, and n can have other subtours.

Gillette [GILL76a] provides a FORTRAN code for the Eastman-Shapiro algorithm. To solve his 10-city sample problem, the code generates 180 subproblems. Gillette reports an IBM 370/168 solution time of 1.79 CPU seconds. Our trial on a Burroughs B6700 required over 51 seconds. If, as a rule of thumb, CPU time increases by a factor of ten for each increase in problem size of ten cities, this method will not be capable of solving even moderately-sized problems of, say, 50 cities in a reasonable time.[6]

4.3.1.2.2 Bellmore-Malone. Constraint (4.4a) is derived from the following theorem:

"The added constraint" $\sum_{i\in S}\sum_{j\in S} x_{ij} \geq 1$, "eliminates all subtour solutions of the k-city problem, as well as all subtour solutions of the (n-k)-city subproblem remaining." [BELL74], where S is any given subtour of length k, and $\bar{S}$ is the complement of S. In a subtour, if one arc in the shortest subtour is changed so that its final node is not in that subtour, the subtour will be eliminated. Observing Figure 4.8 once again, if, for example, we force an arc from city 1 to link with city 4 instead of city 2, then subtour ((1,2),(2,3),(3,1)) cannot exist. Therefore, in this example, if we use the branching shown in Figure 4.10, then subproblems l, m, and n will each have at least one arc in the shortest subtour replaced by an arc from a node in the subtour to a node that is not in the subtour.

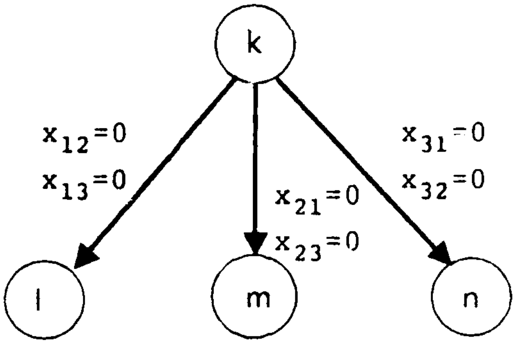


Figure 4.10. Bellmore-Malone branching.

Like Eastman-Shapiro, Bellmore-Malone branching divides the solution space into collectively exhaustive, but not necessarily mutually exclusive, subspaces. Except for subproblems with subtours of cardinality 2, Bellmore-Malone produces more highly constrained subproblems, because it excludes a larger number of arcs. Whereas Eastman-Shapiro prohibits one new arc per subproblem, Bellmore-Malone eliminates $|S| - 1$ arcs, which seems to enumerate solutions by reducing the number of subproblems to fathom.

In a simple computational test, we reprogrammed Gillete's Eastman-Shapiro program to incorporate Bellmore-Malone branching. The revised program generated 119 subproblems, and required only 21.8 CPU seconds. Although we did not run more exhaustive tests (as others have), our conjecture is that this is a representative reduction in computational effort.

Bellmore and Malone offer one additional contribution. For symmetric problems ($c_{ij} = c_{ji} = \forall i,j$), to reduce the occurrence of two-city subtours that hamper the Eastman-Shapiro method, they create mutually exclusive solution subspaces using Murty's assignment ranking technique. [MURT76] Suppose we have the subtour of length k, $x_{i_0 i_1} = x_{i_1 i_2} = \ldots = x_{i_{k-1} i_0} = 1$. Form k subproblems where

$$\begin{aligned} &X_1\colon\ x_{i_0 i_1} = 0; \\ &X_2\colon\ x_{i_0 i_1} = 1;\ x_{i_1 i_2} = 0; \qquad (4.7) \\ &X_3\colon\ x_{i_0 i_1} = x_{i_1 i_2} = 1;\ x_{i_2 i_3} = 0; \end{aligned}$$

and so forth. That is, for the smallest subtour, begin by excluding one arc. In the next subproblem, exclude the next arc in the subtour, and include the arc(s) excluded in the previous subproblem.[7]

In the example in Figure 4.11, subproblems on the right have many arcs excluded. Subproblem n, for example, has 15 exclusions.

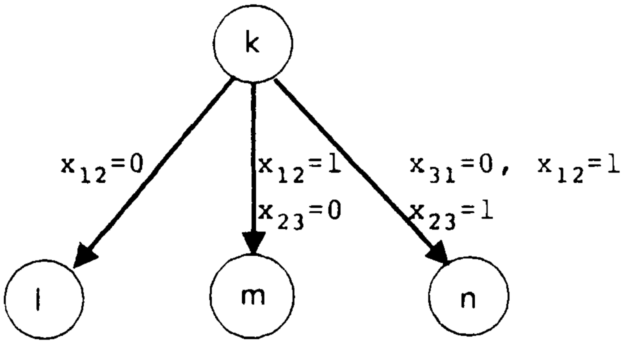


Figure 4.11. Bellmore-Malone-Murty partitioning branching.

	1	2	3	4	5	6
1	M		M	M	M	M
2	M	M		M	M	M
3		M	M			
4		M	M	M		
5		M	M		M	
6		M	M			M

Figure 4.12. Subproblem n arc exclusions.

Not only does this technique increase the number of arcs prohibited, it also leads to the following situation in Gillete's 10-city problem (Figure 4.13). Because we make all arc inclusions and exclusions on all the branches in the path from the root node (subproblem 1) to any given subproblem, the MAP for subnroblem 14 would exclude arc (1,2) because of the branching at node 6, but it would include arc (1,2) in the second branch of node 1. The

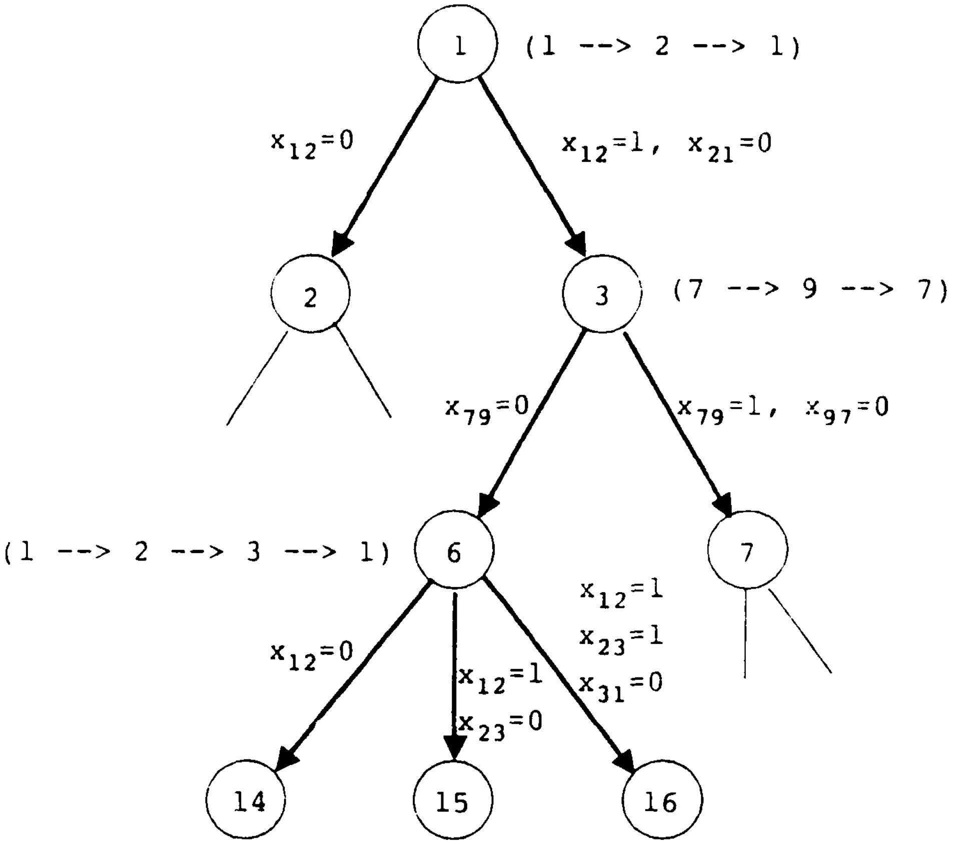


Figure 4.13. Partial solution tree for Gillete's 10-city problem using modified Bellmore-Malone branching.

cumulative effect of including and excluding the same arc is to set all c_{ij} in rows i and j to infinity, which means that the objective value of the solution must be infinite. Therefore, we need not solve subproblem 14. In Gillette's problem, 20 subproblems were fathomed without solving them as the result of this type of conflict.

With the Murty assignment ranking modification, then, the Bellmore-Malone method partitions the solution space at each subproblem node, reduces the number of subproblems that must be solved, and greatly constrains a subproblem through its cumulative arc inclusion feature. Later in our discussion of improvements suggested by Carpento and Toth, we will show that an even higher degree of subproblem constraint can be achieved by choosing a particular subtour for branching, and by selecting the first subtour arc to exclude in a non-arbitrary way.

4.3.1.2.3 Garfinkel. Garfinkel [GARF73] observed that the asymmetric branching method of Bellmore and Malone (the method described above without Murty's modification) does not partition the solution space at each node. If we define

E_k as the set of all arc exclusions on the path from subproblem 1 to subproblem k,

E_k^m as the set of all exclusions for the mth branch of subproblem k,

n as the index of the last subproblem created,

E^{n+i} as the new exclusions for subproblem n+i,

$S_k = \{i_1, i_2, \ldots, i_m\}$ as the set of nodes in the smallest subtour of subproblem k,

$\overline{S_k}$ as the complement of S_k,

then the Bellmore-Malone asymmetric TSP branching rule can be expressed as follows. Create m sets of arc exclusions:

$$E_k^{n+1} = E_k \cup E^{n+1} = E_k \cup \{(i_1,j) \mid j \in S_k \& i_1 \neq j\}$$

$$E_k^{n+2} = E_k \cup E^{n+2} = E_k \cup \{(i_2,j) \mid j \in S_k \& i_2 \neq j\}$$

$$\vdots \qquad (4.8a)$$

$$E_k^{n+m} = E_k \cup E^{n+m} = E_k \cup \{(i_m,j) \mid j \in S_k \& i_m \neq j\}.$$

Construct and solve m MAP's, each with a different arc exclusion set.

Clearly, the Bellmore-Malone method does not partition the solution space. Using our six-city example, if the solution to subproblem 1 was that depicted in Figure 4.8, and we let k = n = 1, then our exclusion sets would be:

$$E_1^2 = \{(1,1),(2,2),(3,3),(4,4),(5,5),(6,6),(1,2),(1,3)\}$$

$$E_1^3 = \{(1,1),(2,2),(3,3),(4,4),(5,5),(6,6),(2,1),(2,3)\}$$

$$E_1^4 = \{(1,1),(2,2),(3,3),(4,4),(5,5),(6,6),(3,1),(3,2)\}$$

Suppose the optimal solution is $x_{15}=x_{52}=x_{26}=x_{63}=x_{34}=x_{41}=1$. Since none of the three arc exclusion sets contains any of the optimal solution arcs, the optimal solution may be in all three solution subspaces created by this branching.

Garfinkel modified this branching strategy so that the solution space is partitioned at every subproblem node. His procedure creates m exclusion sets:

$$E_k^{n+1} = E_k \cup E^{n+1}$$

$$E_k^{n+2} = E_k \cup \{(i_1,j) \mid j \in S_k \text{ bar}\} \cup E^{n+2}$$

$$\vdots \qquad (4.8b)$$

$$E_k^{n+m} = E_k \cup \{\cup_{i=1}\{(i_i,j) \mid j \in S_k \text{ bar}\}\} \cup E^{n+m}$$

The first set is identical to the first created by the Bellmore-Malone method. The i^{th} set contains (1) the cumulative arc exclusion set (E_k), (2) the set of new exclusions (E^{n+1}), and (3), exclusions that prohibit nodes $i_1, i_2, \ldots, i_{i-1}$ from linking with nodes not in the subtour. In other words, as we consider each arc in the shortest subtour, we force some nodes to remain connected with other nodes in the subtour, while forcing the initial node of the subtour arc currently considered to link with a final node outside the subtour. In our example, $E_1{}^3$ contains E^3 $=\{(1,4),(1,5),(1,6)\}$, and $E_1{}^4$ also contains E^4 = $\{E^3,(2,4),(2,5),(2,6)\}$.

To show that the Garfinkel branching rule partitions the set of all solutions, X_k, of subproblem k, we must determine a set of solutions, $X_k{}^* = \{X_k{}^1, X_k{}^2, \ldots, X_k{}^m\}$. $X_k{}^*$ is a partition iff:

$$(ii)\ X_k^i, X_k^j \in X_k^*, i \neq j \textit{ implies } X_k^i \cap X_k^j = \emptyset \ \forall\, i.$$

$$(i)\ \cup_{i=1}^{i=m} X_k^i = X_k, \textit{ where } X_k^i \in X_k^*, i = 1,2,\ldots m\text{: } \textit{and}$$

Any solution to subproblem k will be found in exactly one $X_k{}^i$ if $X_k{}^*$ is a partition. Clearly, $X_k{}^*$ is a partition of x_k, since any solution x' in X_k must have

$$x'_{:tj} = 1 \textit{ for some } t,\ 1 \leq t \leq m \textit{ and } j \in S_k \qquad (4.9)$$

In our example, $x_{1j} = 1$, $j \in S_k$, while $x_{2j} = x_{3j} = 0$ for $j \in S_k$.

Garfinkel did not apparently implement his algorithm, or has not published his results. Tests conducted by Smith, Srinivasan and Thompson [SMIT77] found their version of his algorithm inferior in solution time performance to other methods, including their cost operator technique.

4.3.1.2.4 Srinivasan-Thompson. The Srinivasan-Thompson algorithm [SRIN73] uses their cost operator theory of parametric programming for the well-known transportation problem to accelerate subtour elimination methods for solving Problem TSP. Essentially, their procedure parametrically varies the MAP costs of arcs to be excluded, using the root MAP optimal basis and dual variables, to calculate weak lower bounds for each subproblem, as an alternative to solving subproblems to optimality. Their example uses the Eastman-Shapiro branching method, and they suggest (without further clarification) that weak lower bounding is also compatible with the Bellmore-Malone algorithm.

The Srinivasan-Thompson procedure employs the well-known concept in parametric linear programming that, if the c_{ij} associated with a basic variable x_{ij} in a minimization problem is increased, at some (calculable) limit, c_{ij}', x_{ij} will be replaced in the basis by another (one of possibly several non-basic) variable. Srinivasan and Thompson utilize the fact that, over the interval $(c_{ij}' - c_{ij})$, a series of basis changes involving only degenerate variable exchanges take place. For each exchange, they recalculate affected dual variables and the new objective value,

$$Z_{NEW} = Z_{OLD} + \delta, \qquad (4.10)$$

where δ is the increase in c_{ij} since the last exchange. They call this new value of Z the weak lower bound. If a basis change is made at c_{ij}', x_{ij} will leave the basis, the non-degenerate portion of the solution will change, and the weak lower bound will equal the Eastman-Shapiro lower bound (the subproblem's optimal objective value).

Their method, in essence, generates a sequence of non-decreasing primal feasible solutions, and the process can be halted after any intermediate exchange,[8] after less computation than would be necessary for complete execution of the Hungarian algorithm. By simply recalling the <u>original cost matrix</u>, the optimal dual variables for the root node solution, the <u>complete basis</u> (including degenerate variables), the excluded arc(s) and the latest parametric value of its (their) associated cost(s), the process can be initiated or continued on any subproblem in the solution tree. And, the dual solution and primal basis of any intermediate solution are sufficient to provide a lower bound, thus avoiding having to solve an MAP to optimality.

To demonstrate the algorithm, define

$X = \quad \{x_{ij}\}$, a primal feasible solution of MAP, (i,j), a cell, $i,j=1,2,\ldots,n$, $i \neq j$, and

B = $\{(i,j)\}$, the set of arcs associated with the basis, with $|B| = 2n-1$.

The dual to (4.1)-(4.4) is

$$\textit{Maximize} \quad \sum_{i=1}^{n} u_i = \sum_{j=1}^{n} v_j \tag{4.11}$$

$$\textit{Subject to} \quad u_i + v_j \leq c_{ij} \qquad i,j = 1,2,\ldots,n,\ i \neq j \tag{4.12}$$

$$u_i,\ v_j \ \textit{unrestricted}, \qquad i,j = 1,2,\ldots,n. \tag{4.13}$$

where u_i and v_j are the dual variables associated with constraints (4.2) and (4.3). By the complementary slackness theorem, the 2n-1 equations

$$u_i + v_j = c_{ij},\ (i,j) \in B, \tag{4.14}$$

determine a one-parameter family of dual solutions, and solution $D = \{d_{ij} = u_i + v_j\}$ satisfying (4.11)-(4.13) is a dual feasible solution. Solutions X and D that are both primal and dual feasible are optimal.

By means of an example we will describe how their algorithm works. Suppose for a given subproblem, arc (p,q) is to be excluded. The object of the Srinivasan-Thompson algorithm is to find the maximum amount, μ_i^+, that can be added to c_{pq} to force either a degenerate variable exchange or x_{pq} to leave the basis. For the matrix X (which is the basis of the root subproblem for Gillette's 10-city problem after two degenerate exchanges), with (p,q) = (1,2), their scanning routine identifies chains beginning at (p,q) connecting basic cells already in a chain with another

	1	2	3	4	5	6	7	8	9	10	u_i
1		1	0								56
2	1										-97
3	0	0		1							11
4			1			0					-21
5						1					-10
6				0	1						-4
7			0						1		154
8									0	1	250
9							1				227
10								1	0		248
v_j	287	178	242	210	221	227	-9	87	64	84	

Figure 4.14. Srinivasan-Thompson example basis.

basic cell in the same row or column, much like the Charnes-Cooper stepping stone process. This identifies two sets, I_p = {1,4,5,7,8,9,10}, and J_p = {3,6,7,8,9,10}, that index the rows and columns respectively in chains beginning at (p,q). Sets I_q = {2,3,6} and J_q = {1,2,4,5} are their complements. They then find

$$\mu_i^+ = \min_{(i,j)\ [I_p \times J_q - \{(p,q)\}]} c_{ij} - u_i - v_j, \qquad (4.15)$$

the maximum by which the current parametric value of c_{ij} can be increased, where the candidate values are those of the familiar cell evaluators of the transportation model [MURT76], computed using revised dual variable values. In this instance, the minimum is 6, for (i,j)=(8,2). The previous objective is increased by 6, $x_{8,2}$ is chosen as the entering variable, and a stepping stone procedure determines a leaving variable and makes a degenerate reallocation among the variables in the stepping stone path. The final iterative step updates affected duals by adding $\mu_i{}^{+}$ to u_i's with $i \in I_p$ and subtracting μ from v_j's with $j \in J_q$. When this procedure is executed again, x_{12} is chosen as the leaving variable and the Eastman-Shapiro lower bound of 2500 is reached.

The algorithm offers several advantages over the Eastman-Shapiro strategy. Computations can always be initiated from the root optimal solution for any subproblem, and very little information about a given subproblem needs to be stored. Only simple row and column reduction operations are performed to set up for the scanning and basis change routines. The method relies upon primal transportation solution techniques for which very powerful routines exist to solve the root problem and perform stepping stone and basis change steps, so it should yield considerable computational improvement over Eastman-Shapiro.

The major disadvantage of the Srinivasan-Thompson algorithm is its incompatibility with powerful new bounding techniques [BALA81]. Since they require the optimal subproblem solution, this algorithm would be just a computational alternative, at best decreasing subproblem solution time, but not the number of subproblems solved.

4.3.1.2.5 Carpento-Toth. Carpento and Toth [CARP80b] propose a number of interesting modifications to the Bellmore-Malone approach to the asymmetric TSP, including a

(i) new method for selecting the subproblem to be branched;

(ii) procedure to eliminate unnecessary branchings, and to choose the order of branching non-arbitrarily;

(iii) lower bounding technique to determine if an MAP solution will exceed the current upper bound before reaching MAP optimality; and

(iv) data structuring and other implementation improvements to increase algorithmic efficiency.

These modifications yield considerable improvement over the Smith-Srinivasan-Thompson strategy in tests reported in Management Science. [CARP80b]

Using their definitions, let:

$N =$ $\{1,2,\ldots,n\}$, the vertex set;

$L =$ $\{(i,j) | i,j \in N\}$, the arc set of

$G(N,L)$, a directed graph;

$G(N_1,L_1)$, a graph in which

$N_1 =$ $\{r_1,r_2,\ldots,r_p\}$, $p \leq n$, and

$L_1 =$ $\{(r_1,r_2),(r_2,r_3),\ldots,(r_p,r_1)\}$, is a tour if $p = n$, and a subtour otherwise.

If, at node k, $G(N_1,L_1)$ is a subtour with m arcs, then the Bellmore-Malone method would branch the subproblem into m descendents, where for the j^{th} descendent of k,

$$E_j = E_k \cup \{(r_j, r_{j+1})\}, \text{ and} \tag{4.16}$$

$$I_j = I_k \cup \{(r_\mu, r_{\mu+1}\}) \mid \mu = 1, \ldots, j-1\} \tag{4.17}$$

are its excluded and included arcs respectively.

As pointed out in earlier discussion, if the Bellmore-Malone branching excludes a previously included arc, then we need not solve that subproblem. Carpento and Toth use this property to compute

$$v = e_{\bar{q}} - |L_{\bar{q}} \cap I_k| = \min_{q=1,\ldots,t} \{e_q - |L_q \cap I_k|\}, \tag{4.18}$$

where t is the number of subtours in subproblem k, q indexes the t subtours of subproblem k, L_q is the set of arcs in the q^{th} subtour, I_k is the set of included arcs, and $e_q = |L_q|$. They select the subtour with the smallest v for branching, which is the subtour with the fewest arcs after the number of included arcs is deducted. By using this technique, their algorithm will reduce the number of branchings of individual subproblems, but it is not clear how effective it is in reducing the total number of subproblems solved, which is one of the most important objects of a branching strategy. Since a number of subtours may produce the same v, we presume Carpento and Toth select one arbitrarily when ties occur.

To illustrate, suppose we have the subtour shown in Figure 4.15, with $I_k = \{(4,5),(5,3)\}$ and $E_k = \{(3,4),(5,4),(7,5)\}$. Branches 2 and 3 each exclude an arc in I_k, which means that no feasible solution exists. Therefore, v = 2 indicates that only two branchings are required, even though the subtour contains four arcs.

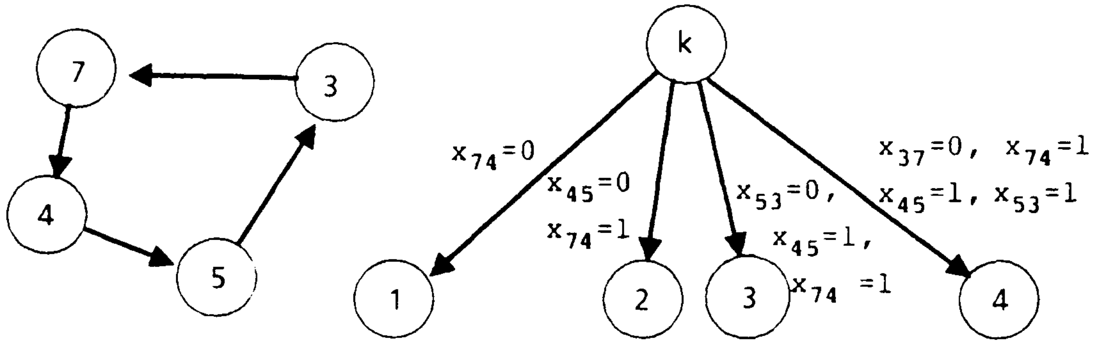


Figure 4.15. Carpento-Toth subtour selection.

In this same example, notice that we achieve the same result if we select (3,7) as the first arc for exclusion. They propose criteria to make this choice in a non-arbitrary way. Let $L_{\bar{q}} = \{I_k - L_q\}$, the set of arcs in the subtour not previously included. Carpento and Toth show that if two included arcs form a path, $\{(k,i),(i,j)\}$, then one additional arc is excluded over including, say, (k,i) and (j,l). If h_j is defined as the number of arcs linked[9] to the jth arc of $L_{\bar{q}}$, then h_j is the number of additional exclusions due to 'not-included' arcs linking with arcs in I_k. They propose the measure

$$w_j = h_j(v-1) + h_{j+1}(v-2) + \ldots + h_v(j-1) + h_1(j-2) + \ldots + h_{j-2}, \tag{4.18}$$

where the parenthetic terms in v and j decrease to zero. They then select arc $(i_u,f_u) \in L_{\bar{q}}$ to exclude on the first branch, where u is such that $w_u = \max\ \{w_j\}$, $j=1,\ldots,v$. In case of ties they use the Little branching variable evaluation technique to choose the first arc. They then generate v branches (omitting those with an arc inclusion/exclusion conflict), with j = v,v-1,...,1 indexing the revised ordering obtained by rotating the arcs in the subtour until (i_u,f_u) is the first arc. Reversing the order ensures that the last branch will be maximally constrained. They also recommend Little's technique of making additional exclusions that prevent subtours among included arcs.

Their third modification, for lower bounding, appears to be the same as Murty's; during the solution of an MAP, if their labeling ends in a non-breakthrough, they compute the lower bound C + kH, where C is the optimal value of the parent subproblem, k is the number of rows without allocations, and H is the minimum element in the labeled rows and unlabeled columns. Murty accumulates reducing constants and the products xH, where x is the difference between the number of labeled rows and unlabeled columns. The subproblem is immediately fathomed if the lower bound (the current dual objective) exceeds the upper bound.

For the fourth improvement, we both use the data structuring technique of maintaining a queue of unfathomed subproblems in non-decreasing order of MAP optimal solution value. Their cost matrix setup technique of finding a common ascendant node for the last subproblem solved and the next one to be solved, removing arc exclusions and inclusions on the path of predecessors of the last problem up to the common node, then inserting the changes on the descendent path to the selected subproblem is more efficient than Gillette's technique of always ascending to the root node. To include arc (i,j), they set all costs except c_{ij} in column j to M, while we also set row i elements to M, which more tightly constrains the MAP.

Our experience and their published results indicate that these independently discovered but virtually equivalent techniques do improve the Bellmore-Malone method substantially. In the next section, we will examine bounding methods that further improve the Bellmore-Malone method.

4.3.1.3 Improved Bounding through Lagrangean Relaxation. Christofides and Balas [BALA81] have recently published new bounding techniques that substantially improve the performance of the Bellmore-Malone algorithm for asymmetric traveling salesman problems.[10] Although they address only the classic TSP case, we will show in the next chapter that their methods improve multiple vehicle and depot subproblem solutions. In this section, we will only describe their algorithm; for complete coverage of these methods, and for proof of supporting propositions, the reader should refer to Reference [BALA81].

The fundamental motivation for developing improved bounding procedures is that

> ... for all branch and bound methods, the quality of the computed bounds has a much greater influence

> on the effectiveness of the algorithm than any branching rules that may be used to generate the subproblems during the search. [CHRI79]

Bounding involves two separate problems: upper bounding, in which the principal concern is finding feasible (tour) solutions; and lower bounding, which seeks better estimates of the optimal objective. The manner in which the two are carried out is completely different.

Since the upper bound is only changed when a feasible solution is found that is better than the incumbent, upper bounding procedures must not only (i) be able to find feasible solutions, but (ii) find such solutions as close to the optimal as possible. Usually, this is done passively; if a subproblem solution is a tour, then it is compared with the incumbent. No process comparable to AP relaxation exists for generating feasible solutions. Some algorithms (e.g., Little et al., and Svestka and Huckfeldt) do attempt to convert subtour solutions into tours, or find a feasible tour through restricted branching, at least for the first subproblem. However, even if a reasonably efficient procedure is used to generate feasible tours, close proximity to the optimal is not assured.

Christofides and Balas propose a method not only for generating a feasible tour for every subproblem; they also attempt to improve those tours if they are not optimal. Since their method is closely tied to lower bounding procedures, we will first introduce those procedures. We will then discuss their upper bounding technique, which generates a tour using reduced costs revised through lower bounding. We will conclude our discussion of their method by mentioning some additional techniques made possible by their bounding methods that greatly reduce the number and size of subproblems to be solved.

In the case of lower bounds, two major classes of relaxations are used to generate lower bounds. The spanning tree approach, credited to Held and Karp [HELD70], exploits the fact that in a network with n nodes and n(n-1) arcs (a completely connected graph without self-loops), a spanning tree is a collection of arcs chosen such that every node can be reached from every other node. Since the optimal TSP tour is the shortest Hamiltonian chain with n arcs, deleting the largest arc in the optimal tour creates an n-1 arc hamiltonian path that is also a spanning tree. The difference between a spanning tree and a hamiltonian path is that the spanning tree does not necessarily consist of a single chain that reaches all n nodes. (In a spanning tree, the nodes do not have to be of degree 2). The shortest spanning tree (denoted SST) will be no longer than the hamiltonian path derived from the optimal tour, and, therefore, the SST will provide a lower bound on the optimal tour.

In the symmetric cost TSP, the methods of Prim [PRIM57] and Kruskal [KRUS56] efficiently find the SST. However, the asymmetric case is computationally more difficult, with the result that each directed SST (DSST) solution on a directed TSP graph requires 4-6 times more effort to solve than a symmetric problem of equivalent size for problems with up to 100 nodes. [CHRI79] Further, the quality of the DSST bound is appreciably inferior to the quality of the SST bound for the corresponding symmetric problem. [CHRI76]

The assignment problem, on the other hand, can be solved efficiently, regardless of whether or not the cost matrix is symmetric. More efficient methods are available for certain cases (e.g., Edmonds matching algorithm for symmetric problems), but the critical point is that the assignment algorithm need not necessarily be changed. Christofides [CHRI79] found the AP solution inferior to the DSST solution, but with additional bounding techniques, Christofides and Balas improve an initial AP bound for each subproblem until it is as close to the optimal as the DSST bound, but requires 10 to 20 times less computational effort. Experience to date indicates that the AP approach, with modifications we will discuss next, is superior to the DSST approach to lower bounding in computational effort, while yielding comparably tight lower bounds.

The essence of the Christofides-Balas method is to start with an AP solution, then solve a restricted Lagrangean problem in which violated subtour prevention constraints are included in the objective via Lagrange multipliers. After applying three bound improvement techniques, they search for a tour, which, if found, is either optimal or provides a new upper bound if better than the incumbent. If the tour is not optimal, additional improvement is sought in the lower bound through use of three additional bounding techniques. If no tour is found, a heuristic is employed to generate one.

They begin with the graph $G_0 = \{N, A_0\}$, where $A_0 = \{(i,j) | \bar{c}_{ij}=0\}$, and the $\bar{c}_{ij}$ are the final reduced costs generated by the AP algorithm. A_0 is called the <u>admissibles matrix</u>, and any tour must consist only of arcs in A_0. In the admissibles graph we will find a tour iff every node is accessible from every other node. We might see, for example, the admissible arc graph shown in Figure 4.16,

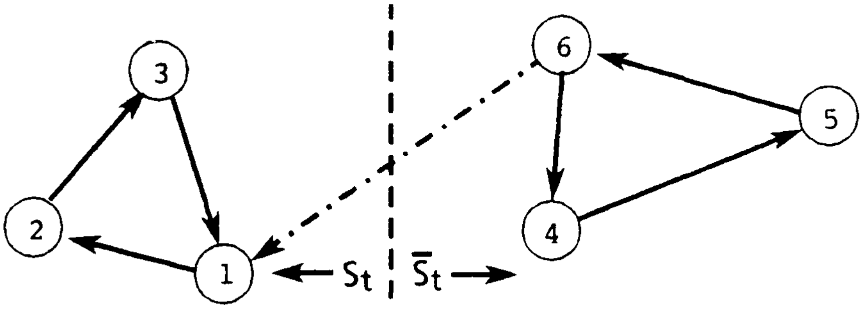


Figure 4.16. Graph G_0.

where the optimal AP solution consists of subtours ((1,2),(2,3),(3,1)) and ((4,5),(5,6),(6,4)). Arc (6,1) is an admissible arc ($c_{6,1} = 0$) not in the solution. An optimal tour consists of arcs from G_0, but a tour cannot exist if we can find a <u>cutset</u> $K_t = (S_t, \bar{S}_t)$, where $\emptyset \in S_t \subseteq N$, for which no arc (i,j) exists in G_0 when $i \epsilon S_t$ and $j \epsilon \bar{S}_t$. We search for the cutset by finding all nodes accessible to node 1, then all those accessible to node 2, etc., until (i), all nodes are accessible from a given node, or (ii), the accessible set of nodes for that node is less than N. If the latter case occurs, then S_t contains the given node plus all arcs accessible from it along arcs in G_0. In the example above, node 1 can reach only itself, node 2, and node 3. Therefore, $K_1 = (S_1, \bar{S}_1) = (\{1,2,3\},\{4,5,6\})$. Therefore, (4.4a) cannot be satisfied without, as a minimum, admitting the arc $(i,j) \in K_1$ corresponding to

$$L_1^0 = min\, \{\bar{c}_{ij}\},\ (i,j) \varepsilon K_1 \qquad (4.19)$$

and penalizing every arc of K_1 by setting $\bar{c}_{ij} = \bar{c}_{ij} - L_1{}^0$, $\forall$ $(i,j) \epsilon K_1$. If v(MAP) is the objective value of the AP optimal, then Christofides and Balas show that a better lower bound is

$$B_1 = v(AP) + \sum_{t \varepsilon T_1} L_1^0 \qquad (4.20)$$

where T_1 is the set of subtours violating (4.4a). If arcs are added to G_0 whose reduced costs $\bar{c}_{ij}$ are reduced to zero, then the resulting graph will be strongly connected when all nodes are accessible from every other node. If, for example, arc (3,6) is added to G_0, then all nodes will be reachable from every node. They prove that, at most, 1/2(h-1)(h-2) cutsets are required, where h is the number of subtours, and since the maximum h is the integer portion of N/2, the cutset procedure is polynomial.

Their second procedure treats violated constraints of the type (4.4b). It utilizes the concept of a cover, a set of row and column indices of the reduced cost matrix such that, if lines were drawn through those rows and columns, all reduced costs of zero in a subtour would be 'lined out'. The procedure finds covering lines, for one subtour at a time, that, (i) cover an allocated arc with both nodes in the subtour exactly once, (ii) cover non-allocated admissible arcs with both nodes in the subtour no more than once, and (iii), do not cover any admissible arcs with only one node in the subtour. Costs $\bar{c}_{ij}$ are further reduced by a penalty,

$$\mu_t = min\,\{\bar{c}_{ij}\},\ (i,j)\varepsilon M \qquad (4.21)$$

where M is the set of arcs with exactly one node in the subtour, and arcs covered by two lines with both nodes in the subtour. Since reduced costs must remain at or above zero, no penalty can be applied if one of the candidate costs is already zero.

The second procedure finds new admissible arcs that eliminate violations of (4.4b); (4.4b) requires that only a path (containing one arc less than a hamiltonian circuit through a subset of nodes) formed by the solution connect the nodes in S_t. If A_t is the set of allocations whose arcs have both ends in subtour S_t, then the subtour will not be eliminated unless there are at least two nodes in the subtour with degree greater than 2 in G_0. If this is not the case, then additional admissible arcs must be found using penalties to reduce costs to zero that are incident into and out of adjacent nodes in that subtour.

Suppose graph G_0 consists of the admissible and allocated arcs shown in Figure 4.16 after the first bounding procedure. Consider node 5. Since only arc (4,5) reaches node 5, and only (5,4) leaves, subtour ((4,5),(5,4)) cannot be eliminated until the degree of node 5 in G_0 is increased, while insuring that reduced costs $\bar{c}_{45}$ and $\bar{c}_{54}$ remain zero.

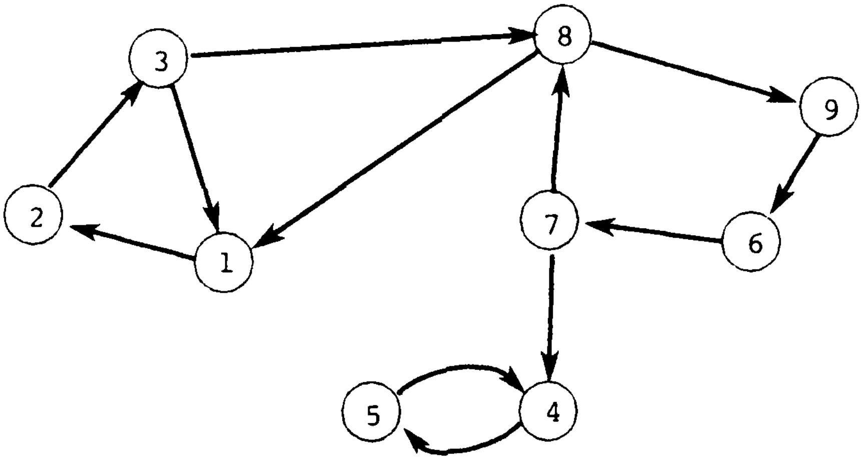


Figure 4.17. Graph G_0 after the first bounding procedure.

The covering procedure utilizes the reduced cost matrix (Figure 4.18) produced by the first bounding procedure. The symbols 0^* represent allocations. The lines enclose the reduced costs of arcs with both nodes in a subtour. In this example, we can find a penalty for subtour ((4,5),(5,4)) by covering row 5 and column 5, selecting the penalty $\bar{c}_{57}$ = 7, and reducing all costs associated with arcs in M before adding the new admissible arc (5,7) to G_0. This procedure is executed once for each subtour.

	1	2	3	4	5	6	7	8	9
1	M	0^*	6	5	9	6	6	5	7
2	4	M	0^*	4	10	6	4	6	9
3	0^*	3	M	5	10	7	3	0	7
4	5	7	9	M	0^*	0	3	7	8
5	4	8	10	0^*	M	10	6	9	7
6	4	4	8	6	11	M	0^*	4	5
7	3	5	4	0	2	9	M	0^*	6
8	0	6	3	5	8	6	9	M	0^*
9	7	3	7	9	6	0^*	5	5	M

Figure 4.18. Reduced costs after first bounding procedure.

Source: [BALA81]

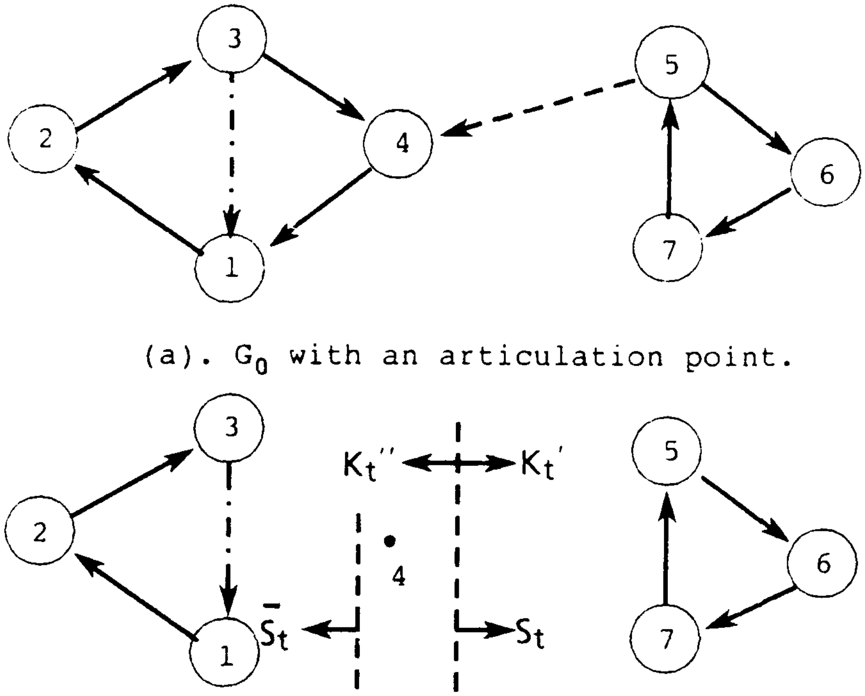
The third bounding procedure eliminates articulation points in G_0. Articulation points are vertices such that, if all arcs incident into and out of them are removed, G_0 is disconnected and at least one of the components is one of the original AP subtours. Any circuit connecting such a subtour would have to include the articulation point at least twice, which means that circuit could not be elementary, which a tour must be. Articulation points are eliminated via a cutset approach.

Figure 4.19a illustrates an admissible graph with an artic-ulation point (node 4). Removing all arcs containing node 4 creates the disconnected graph in Figure 4.19b. Two cutsets are formed: $K_t' = (\{S_t\}, \{\bar{S}_t - \{4\}\})$ and $K_t'' = (\{\bar{S}_t -$

{4}},{S_t}). The penalty

$$v_t = min\{\bar{c}_{ij}\}, \quad (i,j)\varepsilon K_t^{'} \cup K_t^{''} \tag{4.22}$$

is applied to all reduced costs of arcs in both cutsets.



(a). G_0 with an articulation point.

(b). G_0 with articulation point removed.

Figure 4.19. Articulation point removal.

After the third bounding procedure, Christofides and Balas use a multi-path hamiltonian circuit search procedure [CHRI75] to find a tour. Three cases can occur:

1) No tour is found.

2) A tour is found that satisfies all inequalities for which positive multipliers were found in the bounding procedures.

3) A tour is found that violates some of the inequalities in (2).

In the first case, an attempt is made to add arcs in increasing order of reduced cost until a tour can be found. In the second case, the tour is the optimal for the overall problem. In the third case, modified versions of the three bounding procedures are executed, to attempt to satisfy some of the violated constraints, and (iff all violated constraints can be satisfied) find the optimal tour.

As a final improvement, Christofides and Balas recommend that arcs with reduced costs greater than the difference between the lower and upper bound for a subproblem be excluded from the solution (by setting the reduced costs very high). They also provide an alternative branching disjunction, to be used in place of the Bellmore-Malone method under certain circumstances.

4.3.1.4 Comparison of SVRP Methods. Scientifically testing and comparing solution methods is virtually impossible. Intervening variables, such as computer hardware, operating system, and programming language differences; the lack of standardized problems; and variations in programmer skills and techniques, confound comparisons. Data magnitudes can strongly influence solution times. [CARP80b] Computer processing speeds are difficult to compare, forcing us to to use rules of thumb, such as the CDC 7600 is about 3 times as fast as the UNIVAC 1108 [BALA81], and the CDC 6600 is 10 to 50 per cent faster than the UNIVAC 1108. [SMIT77])

Machine storage, algorithm, and program design limitations restrict problem size, so the largest problem solved by a particular implementation may not necessarily indicate an algorithm's maximum potential. And, theoretically proving that an algorithm always finds optimal solutions does not guarantee that result when it is programmed.

Despite these comparability problems, the results of tests described in Table 4.1 strongly suggest that improvements in algorithm design, particularly bounding techniques developed by Balas and Cristofides, have both increased maximum problem size and reduced computation time. What these statistics do not show is that, regardless of the machine used, the algorithms with the best computational times consider far fewer subproblems, which means that the improvement in solution time is because of considerable reduction in work done, and not simply due to the use of faster computers.

4.3.2 Vehicle Routing Problems with Precedence Restrictions and Time Windows. Earlier, we defined a precedence relationship as one in which one entity must be serviced before another. In the last chapter, we described the aeromedical service porblem as the need to transport a patient from an originating to a destination hospital. Obviously, this requires us to visit the origin before the destination, which creates a precedence relationship.

TABLE 4.1

COMPARISON OF ASYMMETRIC SVRP SOLUTION METHODS[a]

Algorithm Developers	Year	Typical Performance		Largest Problem Solved		Host Computer
		Number of Cities	Average Solution Time (cpu secs.)	Number of Cities	Average Solution Time (cpu secs.)	
Eastman	1956					
Shapiro	1966					
Bellmore and Malone	1972	80	165.40[b]	80	165.40[b]	IBM 360/65
Garfinkel	1973	100	59.60[c]	100	59.60[c]	UNIVAC 1108
Smith, Srinivasan and Thompson	1977	100	53.00[c]	180	441.40[e]	UNIVAC 1108
Carpento and Toth	1980	120	16.20[d]	200	35.70[d]	CDC 6600
Balas and Cristofides	1981	100	.71[e]	325	49.66[e]	CDC 7600

[a] Eastman and Shapiro did not report solution results for their algorithms, according to Bellmore and Malone [BELL71]

[b] Source: [BELL71]

[c] Source: [SMIT77]

[d] Source: [CARP80b]

[e] Source: [BALA81]

However, the precedence relationship is in the order or sequence of visits, and does not explicitly involve time windows. The total time to service a patient includes preparation for travel, ground transportation from the hospital to the airport, air transportation, a second ground transportation phase, and stabilization at the destination hospital. [SIVE78] Air transportation time is usually insignificant compared to the other phases, but the other stages are scheduled to conform to the aircraft routing schedule. (The schedule is estimated using expected flight and ground times for the routing sequence.) This means that the orginating hospital is given a time for the patient to board the aircraft, so that aeromedical planners normally do not have to observe pick up time windows. And, there usually are no delivery time windows for patient arrival or service at the destination medical facility. Therefore the aeromedical routing problem is not the more difficult combined routing and scheduling problem.

Obviously, in urgent medical cases, minimizing transportation time can be critically important, but the typical response is to either route a reserve aircraft directly to that patient, or to re-route the closest scheduled aircraft directly to that patient's location. The "in-between" case, the priority category, creates a routing constraint, rather than a time window, since the rules only stipulate pickup within one day. If, in isolated instances, pickup or delivery times must be met, (for example, to accomodate airport operating hours), suitable choice of initial depot departure time will usually suffice, or at worst, re-routing may be required. Again, time windows are not normally part of the aeromedical routing problem, and, therefore, the scheduling problem is avoided.

While precedence constraints and the lack of time windows do not introduce the scheduling complication, precedence relationships do complicate routing problems. Without precedence relationships, three useful properties hold in Problem TSP with symmetric distances. First, if we form the convex hull of cities in the plane, extreme points of the hull will be visited in order. [BARA56] Second, optimal arcs will not cross [FLOO56]. And third, one optimal tour is equivalent in solution value to another constructed by traversing the first in reverse order. Each property gives rise to effective solution methods, particularly heuristics, that will not work with precedence constraints. To see this, note that with the first property, if the clockwise order of extreme points is +1, -1, -2, +2 (where +i indicates the origin of i and -i his destination), visiting the extreme points in order would create routes in which destinations are visited before origins. Psaraftis shows optimal solutions that violate the second property. [PSAR78] The third obviously does not hold, since origin-destination order cannot hold in both directions. The point, then, is that TSP methods cannot solve routing problems with precedence relationships without modification.

Two modifications that observe precedence relationships will be formally introduced in the next section. The first, due to Gavish and Srikanth [GAVI79], introduces additional flow variables into the open tour TSP problem in which constraints can be written that strictly require a patient's origin to be visited before his destination. We will introduce a second approach which requires no additional variables, and with only slight modifications to the Bellmore-Malone branching technique, solves the SVMRP by branch and bound.

4.4 Mathematical Formulation of the Single Vehicle, Many-to-Many Routing Problem (SVMRP). Several problems, including the aeromedical and Dial-a-Ride problems, belong to the general class known as many-to-many routing and sequencing problems.[11] The term many-to-many refers to those problems which have both multiple collection (pickup) and multiple distribution (delivery) points (cities). Many-to-many routing problems occur in a number of situations. To illustrate, suppose we have the requirement to pick up a person at one (origin) node, and deliver that person to one (and only one) destination node. Thus, instead of having a single source generating trips to many destinations, or many origins generating trips to a single destination, as in the TSP, we have many origins generating trips to many destinations, hence the term many-to-many.

When the origin-destination precedence relationship is added to Problem TSP, the order of visits to cities implies, correctly, that an origin be visited before the destination for each person. Stated differently, we must find ordered sequences of n arcs

$$O = \{((i_0,i_1),(i_1,i_2),\ldots,(i_n,i_0))\}, \qquad (4.23)$$

where, if i_a is the origin of a passenger, and i_b is his destination, then $a < b$, for every passenger. Of course, sequences that observe this ordering must also be tours. This ordering restriction is the principal difference between many-to-many and TSP problems.

In the following we assume:

(i) Each node (except the depot) is exclusively either an origin or destination for one and only one passenger.[12]

(ii) The depot is neither an origin nor a destination for any passenger.

(iii) There are n passengers, and hence 2n + 1 nodes.

We will label the origin and destination nodes using two different conventions. In the first, let the

$$\text{node label of } j = \begin{cases} i & \text{if } j \text{ is the origin of patient } i \\ n+1 & \text{if } j \text{ is the destination of patient } i \\ 0 & \text{if node } j \text{ is the depot} \end{cases}$$

In the second scheme, let

$$\text{node label of } j = \begin{cases} +i & \text{if } j \text{ is the origin of patient } i \\ -i & \text{if } j \text{ is the destination of patient } i \\ 0 & \text{if node } j \text{ is the depot} \end{cases}$$

The first scheme facilitates the use of summations in mathematical expressions, while the second is useful in graphical presentations and algorithm development.

Gavish and Srikanth [GAVI79] formulate the SVMRP as an integer linear program. In their formulation, they split the central depot into two nodes, labeled 0 and 2n+1, and add the restriction that the vehicle must initially depart node 0 and must terminate at node 2n+1, never arriving at node 0 nor departing node 2n+1. Thus, feasible solutions must be Hamiltonian paths, but not circuits.[13]

The Gavish-Srikanth formulation is to

(P1) Find variables x_{ij}, y_{ij}, i,j=1,2,...,2n+1, that

$$Minimize\ Z = \sum_{i=0}^{2n+1} \sum_{j=0}^{2n+1} c_{ij} x_{ij} \tag{4.24}$$

$$subject\ to: \quad \sum_{i=0}^{2n+1} x_{ij} = 1 \qquad j = 1,2,\ldots,2n+1; \tag{4.25}$$

$$\sum_{j=0}^{2n+1} x_{ij} = 1 \qquad i = 1,2,\ldots,2n+1; \tag{4.26}$$

$$\sum_{j=0}^{2n+1} y_{ij} - \sum_{j=0}^{2n+1} y_{ji} = 1 \qquad i = 1,2,\ldots,2n+1; \tag{4.27}$$

$$y_{ij} \le (2n+1) x_{ij} \qquad i,j = 1,2,\ldots,2n+1; \tag{4.28}$$

$$\sum_{j=0}^{2n+1} x_{j0} - \sum_{j=0}^{2n+1} x_{2n+1,j} = 1 \tag{4.29}$$

$$x_{ij} = 0,1;\ y_{ij} \ge 0;\ \forall\ i,j \tag{4.30}$$

$$\sum_{j=0}^{2n+1} y_{n+i,j} - \sum_{j=0}^{2n+1} y_{ij} = 1 \qquad i = 1,2,\ldots,n. \tag{4.31}$$

They establish the following conditions that must be met by a solution to the many-to-many problem:

(i) the vehicle must arrive at a non-terminal node once and only once,

(ii) the vehicle must depart a non-terminal node once and only once,

(iii) the vehicle must depart node 0 once and only once, and never return to it,

(iv) the vehicle must arrive at node 2n+1 once and never depart it,

(v) the route must not contain subtours, and

(vi) the origin must be visited before its associated destination for each patient.

They prove that P1 meets these conditions:

1) Constraint (4.25), for j=1,2,...,2n, satisfies (i), and constraint (4.26), for i=1,2,...,2n, satisfies (ii). Together, these constraints insure one visit to every non-terminal node.

2) Constraint (4.26), for i=0, and constraint (4.29) ensure that the vehicle departs but does not arrive at node 0, satisfying (iii).

3) Constraint (4.25), for j = 2n+1, and constraint (4.29) ensure that the vehicle arrives at, but does not depart, node 2n+1.

4) Constraint (4.27) ensures that subtours do not occur. Observe that subtours cannot contain node 2n+1 or node 0, since constraint (4.29) restricts the degree of those nodes to 1. Suppose we had the subtour shown in Fig. 4.20.

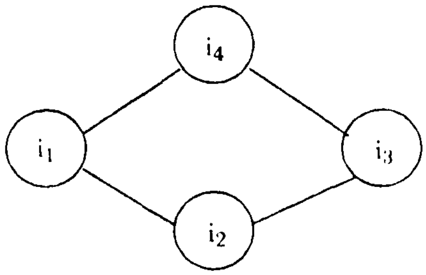


Figure 4.20 Four-city subtour.

Let $y_{i_1 i_2} = k$. From constraints (4.25-4.27),

$$\sum_{j=0}^{2n+1} y_{i_2 j} - \sum_{i=0}^{2n+1} = \sum_{j=0}^{2n+1} y_{i_2 j} - k = 1,$$

$y_{i_2 i_3} = k+1$. Similarly, $y_{i_3 i_4} = k+2$ and $y_{i_4 i_1} = k+3$. Then, for node i_1:

$$\sum_{j=0}^{2n+1} y_{i_1 j} - \sum_{j=0}^{2n+1} y_{j i_1} = k - (k+3) = -3,$$

which violates constraint (4.31). Therefore, constraints (4.29) and (4.27) satisfy (v).

5) Constraints (4.31) and (4.27) satisfy (vi). Constraint (4.27) forces a strict unit increase in flow in successive arcs in the solution path. The first summation term of constraint (4.31) is the flow out of the destination for passenger i. The second summation is the flow out of the passenger's origin. By forcing the first flow to exceed the second by at least one unit, the destination node must be at least the first node after its corresponding origin node in the solution path.

Gavish and Srikanth also note that constraint (4.29) is unnecessary if $c_{i0} = c_{0,2n+i} = \infty \; \forall \; i$. They do not let $c_{n+i,i} = \infty \; \forall \; 1 \le i \le 2n$. Logically, the directed arc from a destination to its corresponding origin should be prohibited, eliminating 2n variables. They also do not let $c_{i,2n+1} = \infty$, $i \le n$, even though the vehicle should not be routed from an origin directly to the terminal.

Gavish and Srikanth do not explain the two functions of constraint (4.28). First, it serves as an 'either-or' constraint, so that if $x_{ij} = 0$, then $y_{ij} = 0$. Second, since there are 2n+1 arcs in a Hamiltonian path through n nodes, and since x_{ij} is either 0 or 1, (4.28) restricts y_{ij} to a value no larger than 2n+1. To illustrate, suppose n = 3. Figure 4.21 shows one possible path.

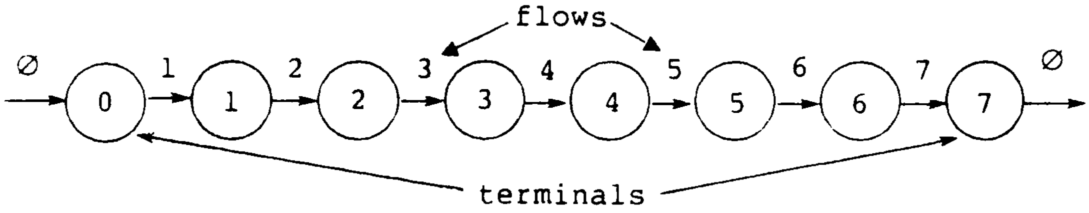


Figure 4.21. Hamiltonian path through 2n+1 nodes.

In this case we observe strict unit increase in flow such that the last arc has a flow of 2n+1 = 2(3)+1 = 7.

This mixed integer formulation (the y_{ij} variables are not required to be integer) requires $2(n^2 - 2n)$ variables and $4n^2 + 8n + 3$ constraints. Since Gavish and Srikanth do not suggest any solution method or computational results, we are not aware of any attempts to directly exploit this formulation. For even small values of n, their formulation would create a very large problem.

Now, consider a second formulation derived from the traveling salesman problem. In this approach we make the same assumptions as Gavish and Srikanth, except that we do not create two terminal nodes. Also, we use slightly different notation in our discussion. Table 4.2 shows the correspondence between the two node labeling schemes. Node 4, the destination of passenger 1 in the first scheme, is labeled node -1 in the second. Node 0 is the terminal. In the following discussion we use these labeling conventions interchangeably.

TABLE 4.2

NODE LABEL EQUIVALENCES

	Node Labels						
Scheme	Terminal	Origins			Destinations		
1	0	1	2	3	4	5	6
2	0	+1	+2	+3	−1	−2	−3

We continue to let c_{ij} be the cost of traveling directly from node i to node j, and $c_{ii} = \infty \forall i$. And, we let $c_{n+i,i} = \infty$, $i \geq 1$, and $c_{i,0} = \infty$, $1 \leq i \leq n$, since the vehicle should not be routed directly from +i to −i (n+i to i) for any passenger, nor should the vehicle go directly from an origin to the terminal, as this would imply at least one passenger not being delivered. Finally, let $c_{0,i} = \infty$, $1 > n$ to prevent the vehicle from going directly from the terminal to a delivery point.

Before we present our formulation, we need to introduce the concept of an *infeasible chain*. To do so, suppose for a seven-city problem, we solved the traveling salesman problem, and obtained the optimal tour in Figure 4.22, where the node labels are from Table 4.2, S is the set of *properly sequenced nodes*, and $\bar{S}$ is one (in this example, the only) set of *improperly sequenced* nodes. We define an *infeasible chain* as a path beginning and ending at a pair of improperly sequenced nodes. In this example, the infeasible chain is ((5,1),(1,2)), using the first labeling scheme. The length of an infeasible chain is the number of arcs between the first and last improperly sequenced nodes.

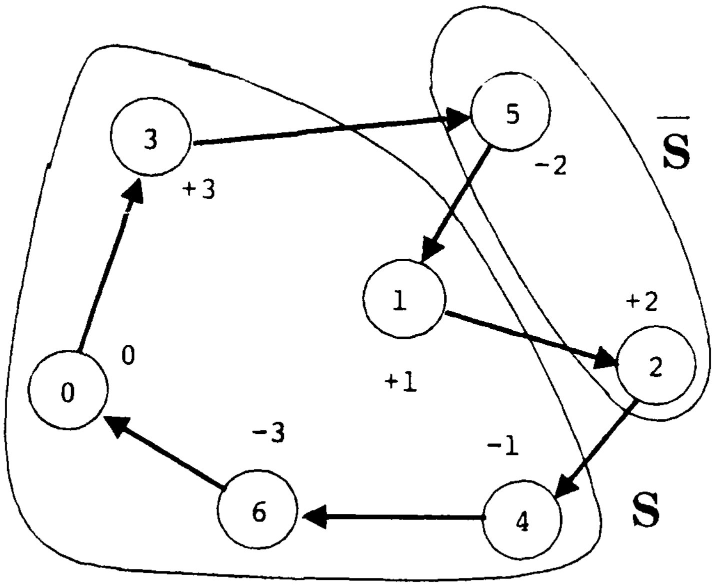


Figure 4.22. An optimal TSP tour.

TSP solution procedures will not eliminate infeasible chains without modification. Even if we find the optimal TSP tour, TSP(x*), if it contains an infeasible chain, then that tour cannot solve the SVMRP. Before we propose an SVMRP solution technique, there are a number of properties we should formally establish about SVMRP solutions and infeasible chains.

Theorem 1. If TSP(x*) is the optimal solution to Problem TSP, and SVMRP(x*) is the optimal solution to Problem SVMRP, then TSP(x*) ≤ SVMRP(x*).

Proof. By definition, TSP(x*) is the shortest tour of n nodes. By definition, any SVMRP solution must be a tour. Therefore, SVMRP(x*) cannot be less than TSP(x*). Q.E.D.

Remark 1. The shortest subtour is of length 2.

Proof. Self-loops are prohibited. By definition, a subtour is a path, and a subtour must therefore have the same initial and final node. At least a second node must be included in the subtour path; otherwise, it would be a self-loop. We can easily construct a subtour of two nodes (see Figure 4.1); therefore, a subtour of length 2 exists. Because subtours of length 1 are impossible, the shortest possible subtour is of length 2. Q.E.D.

Lemma 1. The longest subtour of N cities is N-2.

Proof. If we partition N, and restrict the size of the smallest disjoint subset of N to 2, then the largest disjoint proper subset of N is N-2. The length of a subtour is the number arcs it contains, which by Lemma 1, is also the size of the node set of the graph of the subtour. Therefore, since the largest disjoint proper subset of N is N-2, the length of the largest subtour is also N-2. Q.E.D.

Lemma 2. The longest infeasible chain in a tour of N cities is N-4.

Proof. The length of a tour of N cities is N. Referring to Figure 4.23, representing an N-city tour, if we continue to assume that $c_{01} = \infty$, i > n, then the first arc from node 0 must be incident into a passenger origin. Similarly, the arc incident into node 0 must be incident out of a destination node. Therefore, the first and last nodes of any feasible solution must be properly sequenced. Since we have defined the length of an infeasible chain to be the number of arcs linking two improperly sequenced nodes, then an infeasible chain can begin no earlier than the node adjacent to the first (properly sequenced) node after node 0. The chain must end at least at the second node prior to node 0 at the end of the circuit. Therefore, four arcs cannot be included in a feasible chain. Since we can construct an example of an infeasible chain of length N-4 (Figure 4.22), one exists. Therefore, the longest infeasible chain of N cities is of length N-4. Q.E.D.

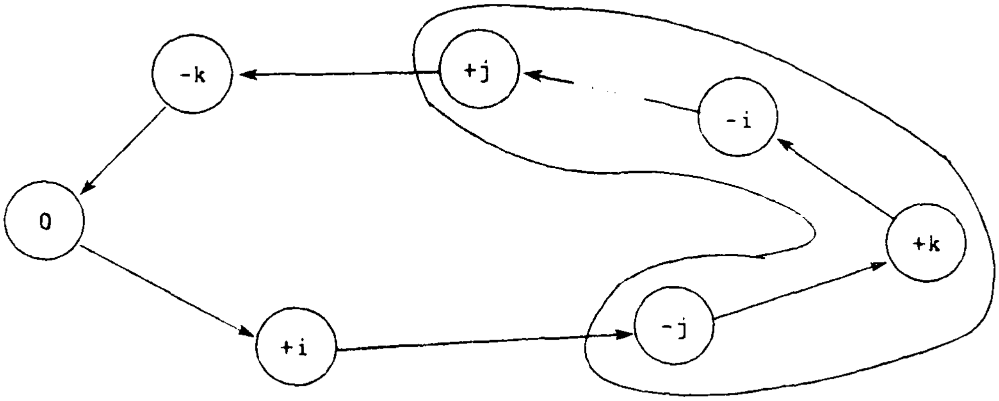


Figure 4.23. An infeasible N-city SVMRP tour.

Lemma 3. The length of the shortest possible infeasible chain in a tour or subtour of N cities is 2.

Proof. By Remark 1, the shortest tour or subtour is of length 2. The minimum length of an infeasible chain is 1, because an infeasible chain must be a path between two improperly sequenced nodes, and a path must contain at least one arc by definition. However, infeasible chains of length 1 are prohibited, because directed arcs from destnations to origins are prohibited. Therefore, the minimum length must be at least 2. We can construct an example of an infeasible chain of length 2, ((-1,+2),(+2,+1)); therefore, the shortest possible length is 2. Q.E.D.

Lemma 4. An infeasible chain cannot occur in a subtour of length less than three, nor in a tour of length less than 7.

Proof. For an infeasible chain to occur in a subtour of length 2, the two nodes would have to be +i and -i. One arc would have to be (-i,+i), by the definition of a subtour. However, that arc is prohibited by definition. Therefore, subtours must be of length 3 or greater to contain an infeasible chain. Q.E.D.

In many-to-many tours, N must be an odd integer, since there must be two cities associated with each passenger, plus the terminal. To show that N must be at least 7, we can eliminate 1, 3, and 5 as possible values of N. N = 1 implies a no-passenger problem. With only one passenger, and N = 3, an infeasible chain cannot exist because, by Lemma 2, its length would be 0. N = 5 is also impossible, because by Lemma 2, the length of the longest infeasible chain would be 1, which contradicts Lemma 3. Figure 4.24

shows a tour where N = 7, with infeasible chain ((-2,+3),(+3,-1),(-1,+2)).

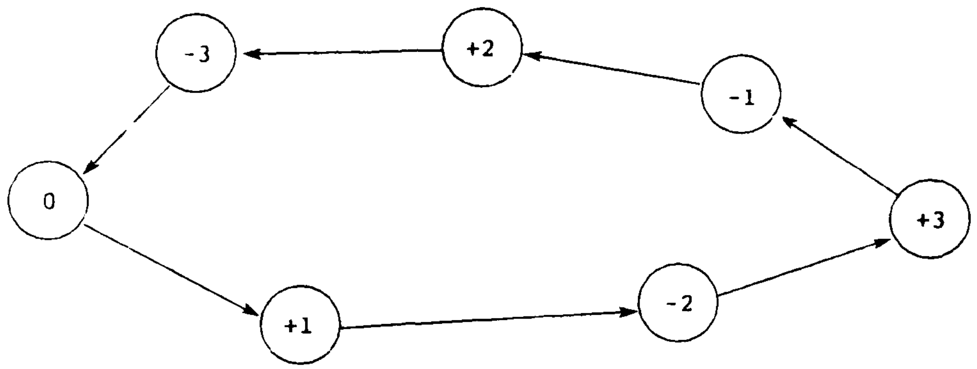


Figure 4.24. An infeasible 7-city SVMRP tour.

That infeasible chains exist with N greater than 7 can be shown by construction. Given an infeasible chain for any N ≥ 7, an infeasible chain for N + 1 can be constructed by simply inserting[14] the two new passenger nodes anywhere in the chain. The sequencing of the previously improperly sequenced nodes cannot be changed by an insertion, so they remain improperly sequenced. Therefore, tours must be of at least length 7 to contain infeasible chains. Q.E.D.

As Gavish and Srikanth have shown, the many-to-many problem is equivalent to Problem TSP with the additional condition that, for every passenger, the solution provides for visiting his origin before his destination. Therefore, if N = 2n+2, where n is the number of passengers, the second formulation of the many-to-many problem is to

Problem SVMRP:

Find variables x_{ij}, i,j = 0,1,2,...,2n that will

$$\text{Minimize } Z = \sum_{i=0}^{2n} \sum_{j=0}^{2n} c_{ij} x_{ij} \tag{4.32}$$

$$\text{subject to: } \sum_{i=0}^{2n} x_{ij} = 1, \qquad j = 0,1,2,\ldots,n; \tag{4.33}$$

$$\sum_{j=0}^{2n} x_{ij} = 1, \qquad i = 0,1,2,\ldots,n; \tag{4.34}$$

$$X = [x_{ij}] \in T \tag{4.35}$$

$$X = [x_{ij}] \in O \tag{4.36}$$

$$x_{ij} = 0 \text{ or } 1, \qquad i,j = 0,1,2,\ldots,2n. \tag{4.37}$$

This formulation contains several differences from Problem P1. Only one terminal is used. Flow variables are not used; instead, constraint (4.36), where O is defined by (4.23), restricts solutions to those which contain proper sequences of nodes. In the following sections, we will examine a number of alternatives for solving Problem SVMRP, and then we will present a modification to the Bellmore-Malone method that will solve (4.32)-(4.37).

4.4.1 Single Vehicle, Many-to-Many Solution Methods. In the next three sections, we will examine three different SVMRP solution techniques. Although in each instance we will find a major problem applying these technique to the aeromedical problem, each has interesting features that could potentially be incorporated in our solution model, which we propose as a promising area for future research.

4.4.1.1 Dynamic Programming. Psaraftis [PSAR78] has devised an exact, dynamic programming (DP) approach to SVMRP. His major application has been the so-called Dial-A-Ride Problem, in which small buses or vans are routed to pick up and deliver customers (typically the elderly, handicapped, or indigent, in urban areas) who telephone a dispatcher for service. He treats two versions, the subscriber case, in which all customer demands are known before routes are determined, and the demand-responsive case, where customer requests received after vans are dispatched are handled by dynamically revising routes. The dial-a-ride and aeromedical problems are very similar, but differences do exist, such as the use of priorities in the aeromedical case.

Assuming that each customer has a unique origin and destination, Psaraftis's model finds an optimal open tour[15] subject to certain constraints. Let

N = the number of customers,

$2N+1$ = the number of nodes, including the depot,

t_{ij} = the travel time between nodes i and j,

T_j = the duration of the jth leg of a route, $T_j \in T = \{t_{ij}\}$, $j = 1,2,...,2N$,

WT_i = the waiting time, from $t = 0$, the vehicle's departure time from the depot, until customer i boards, and

RT_i = the riding time of customer i.

Psaraftis' model incorporates the linear weighted objective

$$Minimize\, Z = w_1 \sum_{j=1}^{2N} T_j + w_2 \sum_{i=1}^{N} (\alpha\, WT_i + (2-\alpha)\, RT_i) \qquad (4.38)$$

which contains two objectives, minimizing total vehicle travel time to service all customers, and minimizing total negative utility of customer waiting and riding.

His model incorporates three major constraints. Routes must observe origin/destination ordering, and the route must not have subtours. Vehicle capacity is explicitly observed. If customers are arranged in a list according to the time of their service request, then, in the sequence of pickups and in the sequence of deliveries, each customer must appear in each sequence at a sequence position (SP) within the range LP-MPS ≤ SP ≤ LP+MPS, where LP is his list position, and MPS is a maximum allowable shift in sequence position. This constraint stems not from a desire to treat customers fairly; rather, it is a necessary device to prevent the DP model from indefinitely deferring a customer because the travel time to serve him is high.

Psaraftis provides an excellent discussion of the combinatorics of Problem SVMRP, and of implementation issues. His algorithm, though exponential ($(O(n^2 2^N))$), provides very reasonable solution times for problems with N ≤ 8, a limit imposed by computer core storage limitations. However, this size limitation and the current lack of multiple vehicle capability render the model incapable of solving Versions III and IV of the aeromedical model. In the future, research should be undertaken to incorporate the useful features of the model (composite objective and ability to handle additional constraints) in a multiple vehicle framework. Also, his techniques for dynamically altering routes to accommodate requests received after a vehicle begins its route are potentially useful in aeromedical airlift mission management, to handle priority and urgent cases dynamically, a problem we do not treat in this thesis.

4.4.1.2 Mixed Integer Programming. Sexton [SEXT79] includes a new service requirement in Problem SVMRP: each customer specifies a desired delivery time. Assuming no constraint on customer pickup time, he defines two causes of customer disutility. The first is excess ride time (ERT), the difference between the time actually spent in the vehicle and the travel time from his origin directly to his destination. Sexton defines delivery deviation time (DDT), the difference between actual and desired delivery time, as the second. If ERT + DDT is the total inconvenience of a customer (TIC), then his objective is to minimize total inconvenience of schedule (TIS), the linear sum of all TIC's.

Unlike Psaraftis, Sexton does not consider waiting time or vehicle capacity. His solution provides both the order of service (route) and times of service for both pickup and delivery for every customer (schedule). His heuristic method uses two exact methods as subroutines, a special case of the transportation model for routing, and Benders mixed integer programming for scheduling. His major application is also the dial-a-ride problem.

In our formulation of the aeromedical problem we have ignored delivery time as a service requirement, which negates the value and applicability of Sexton's model. However, time windows, such as airfield operating hours and time limits on service availability (refueling, for example) can arise, and Sexton's techniques are designed to handle them. We recommend this for future investigation.

4.4.1.3 Other Methods. Because Problem SVMRP is directly related to Problem TSP, which is NP-complete, no efficient solution procedure is likely to be found for it. For that reason, in order to provide any solution at all, let alone an optimal one, researchers have paid increased attention to heuristics. Psaraftis proposes three such procedures.

In each, he uses a simple total distance objective. His goal is to develop a technique that, in its 'worst-case behavior', comes within some acceptable range of the optimal, say $(|L^h - L^*|)/L^* \leq K$, where L^h and L^* are the heuristically detemined and optimal path lengths respectively, and K is a multiple of L^*.

The first procedure [PSAR81] removes infeasible arcs on an unrestricted TSP tour:

Heuristic 1:

Step 1. Solve Christofides' (or any other tour-generating) algorithm for 2N points. Let T be the tour.

Step 2. Pick any point in T as the starting point. Construct a new path T_i' as follows. Moving clockwise, visit each node in T. If the node is legitimate (any pickup point, or any delivery point whose origin is already in T_i'), connect it to the last legitimate node. Continue until all nodes are connected.

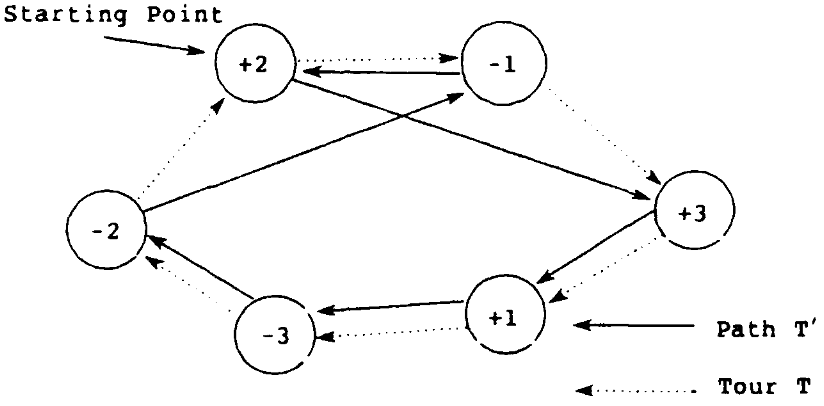


Figure 4.25. Step 2: constructing T′.

Step 3. Repeat Step 2, for all starting points i, and then proceed counter-clockwise to construct T_i''.

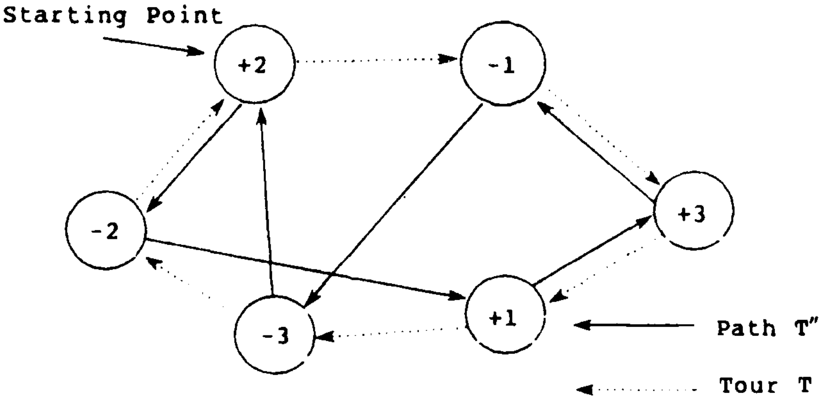


Figure 4.26. Step 3: constructing T″.

Step 4. Select $T^* = \min_{i=1,2,\ldots n} \{T_i', T_i''\}$.

His second algorithm [PSAR81], similar to one proposed by Stein [STEI78], is based on the property due to Barachet [BARA57] that TSP tours visit extreme points in order.

Heuristic 2:

Step 1. Form two tours, one with origins only, and one with destinations only.

Step 2. Connect the two graphs to form one tour.

The worst case behavior of the second method is K = 3. The third procedure, which he calls tree circumnavigaticn, uses the TSP property that a TSP tour is a spanning tree with one additional arc. [PSAR82a]

Heuristic 3:

Step 1. Form the minimal spanning tree.

Step 2. Replace each undirected arc between two given nodes with two directed arcs in opposite directions between those nodes.

Step 3. Find shortcuts while circumnavigating the tree to construct the tour T:

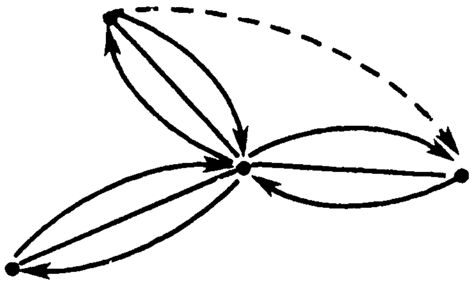


Figure 4.27. Circumnavigation shortcut.

Step 4. Use Heuristic 1 to find T'.

Step 5. Search for arc changes to improve the solution, using a method such as Lin's k-opt technique [LIN 73].

In terms of performance, these heuristics have been very promising in initial tests. With reasonable accuracy and processing time, and procedural simplicity, these methods offer excellent means for constructing initial upper bounds. With further development and refinement, these techniques should be the subject of additional research to determine how well they can accelerate optimization algorithms.

4.5 A Single Vehicle Routing Algorithm for the Many-to-Many Case. In this section we will propose an algorithm that solves Problem SVMRP with a modified version of the Bellmore-Malone-Murty algorithm. We will first demonstrate that this algorithm is superior to modified versions of the Eastman-Shapiro and Garfinkel techniques. After presenting the algorithm, we will demonstrate its use in solving a Version I aeromedical routing problem.

4.5.1 Infeasible Chain Elimination. In the mathematical formulation of Problem SVMRP, we defined an infeasible chain as a path connecting two improperly sequenced origin and destination nodes. This infeasibility can occur in a tour of $N \geq 7$ cities, or a subtour of $|S| \geq 3$ cities. Figure 4.28 illustrates a tour with an infeasible chain.

TSP fathoming tests will fathom a subproblem with a tour solution containing infeasible chains. Therefore, we need to modify TSP fathoming and branching rules in such a

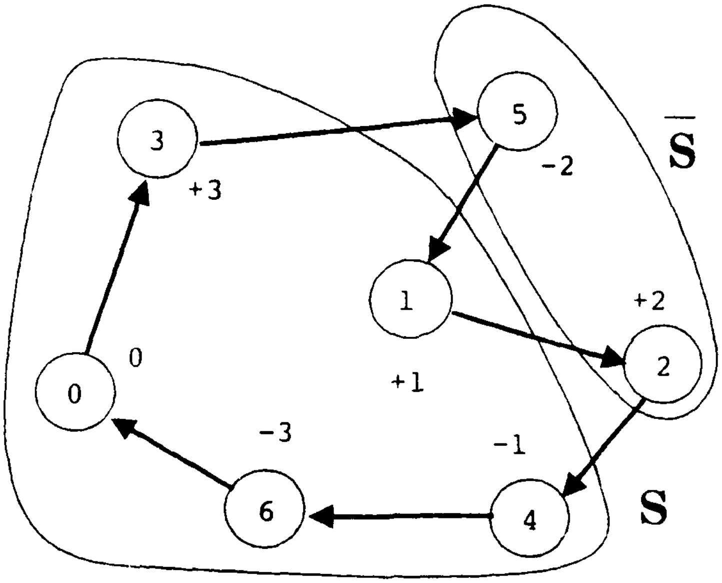


Figure 4.28. TSP tour with an infeasible chain.

way that (i) infeasible chains are eliminated, and (ii) the solution space is partitioned. The tour fathoming test can be easily changeo to detect infeasible chains, and all of the branching methods described above can be modified to eliminate them. We will show, however, that only one will partition the solution space, and provide the most tightly constrained subproblems.

An Eastman-Shapiro strategy, in which we prohibit a different arc of the infeasible chain on each branch, will produce the branching shown in Fig. 4.29 for our example above. Clearly, both branches will eliminate the infeasible chain, but will not partition the solution space. For example, $x_{02} = x_{23} = x_{31} = x_{15} = x_{54} = x_{46} = x_{60} = 1$ can be in both solution spaces (assuming predecessor branchings do not prohibit any of the optimal arcs).

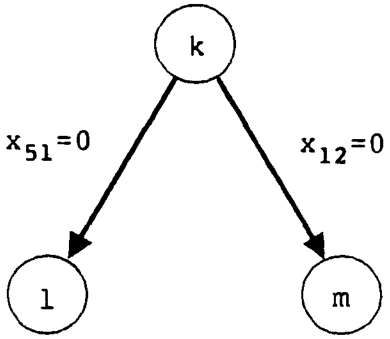


Figure 4.29. Eastman-Shapiro infeasible chain elimination.

If we use Bellmore-Malone-Murty branching, we will partition the solution space. To see this, note the two assignment matrices that result. Subproblem l has the same arc exclusion for both the Eastman-Shapiro and Bellmore-Malone-Murty methods, while the additional exclusions in the second column and sixth row due to the arc inclusion x_{51} = 1 would not ue made by Eastman-Shapiro branching. Subproblem l excludes any solution with x_{51} = 1, while subproblem m excludes all solutions except those with x_{51} = 1.

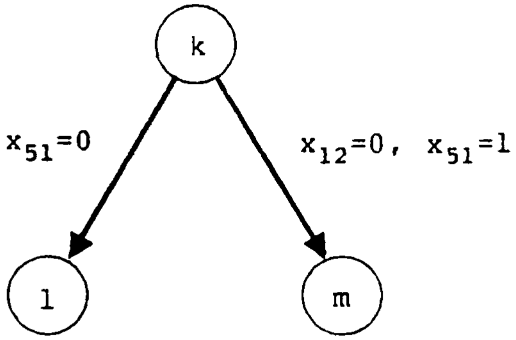


Figure 4.30. Bellmore-Malone-Murty infeasible chain elimination.

$$\begin{array}{c|ccccccc|}
 & 0 & +1 & +2 & +3 & -1 & -2 & -3 \\
\hline
0 & M & & & & M & M & M \\
+1 & M & M & & & & & \\
+2 & M & & M & & & & \\
+3 & M & & & M & & & \\
-1 & & M & & & M & & \\
-2 & & \infty & M & & & M & \\
-3 & & & & M & & & M \\
\hline
\end{array}$$

(a). Subproblem 1 arc exclusions.

$$\begin{array}{c|ccccccc|}
 & 0 & +1 & +2 & +3 & -1 & -2 & -3 \\
\hline
0 & M & \infty & & & M & M & M \\
+1 & M & M & \infty & & & & \\
+2 & M & \infty & M & & & & \\
+3 & M & \infty & & M & & & \\
-1 & & M & & & M & & \\
-2 & \infty & & M & \infty & \infty & M & \infty \\
-3 & & \infty & & M & & & M \\
\hline
\end{array}$$

(b). Subproblem m arc exclusions.

Figure 4.31. Bellmore-Malone-Murty arc exclusions.

Garfinkel's method partitions the solution space, but contains fewer arc exclusions overall because it does not use arc inclusions. Again note the resulting assignment matrices for the two branches. The two branches either require or prohibit solutions with either X_{52} or $x_{53} = 1$. Comparing the Garfinkel strategy to the Bellmore-Malone-Murty approach, we observe that although both partition X_k, the latter introduces more exclusions. Because of this, in

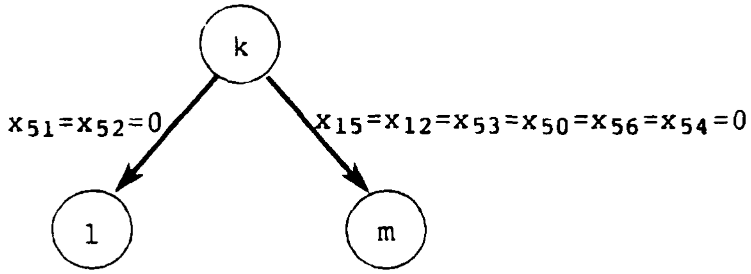


Figure 4.32. Garfinkel infeasible chain elimination.

the algorithm to be presented in the next section, we will use the Bellmore-Malone-Murty strategy for eliminating infeasible chains.

	0	+1	+2	+3	-1	-2	-3
0	M				M	M	M
+1	M	M					
+2	M		M				
+3	M			M			
-1		M			M		
-2		∞	M			M	
-3				M			M

(a). Subproblem l arc exclusions.

	0	+1	+2	+3	-1	-2	-3
0	M				M	M	M
+1	M	M	∞			∞	
+2	M		M				
+3	M			M			
-1		M			M		
-2	∞		M	∞	∞	M	∞
-3				M			M

(b). Subproblem m arc exclusions.

Figure 4.33. Arc exclusions generated by Garfinkel's method.

One aspect of infeasible chains should be mentioned. In the fathoming test, the branching rule is to branch on the shortest subtour or infeasible chain, whichever is detected first, so that fewer branchings will be required for each unfathomed subproblem. Of course, this does not guarantee that fewer subproblems will be solved overall.

4.3.2 The SVMRP Algorithm. In the SVMRP algorithm, let:

B = the number of patients;

N = the number of cities including the depot;

C = the cost matrix $[x_{ij}]$;

O = the order vector where element i = +j if city i is the origin of patient j, -j if city i is the destination of j, and 0 if i is the depot.

CHECK = order feasibility vector where element i = 1 iff node i is correctly ordered, and 0 otherwise.

BFSD = best feasible solution currently known;

CLUB = current least upper bound corresponding to BFSD;

BFSD* = the optimal solution vector;

CLUB* = the optimal solution objective value;

P = the index of the last subproblem solved;

k = the index of the subproblem selected for further branching;

AS(k) = the solution of assignment subproblem k;

Z_k = the cost associated with AS(k);

S(i,j) = the jth successor of node i. (If i=1, and j=2 in the path ((1,2),(2,3)), S(1,2)=3.

CLLB* = $\min_{k=1,...,P} \{Z_k\}$;

Q = the queue containing active (unfathomed and unbranched) subproblems in order of non-decreasing values Z_k;

E_j = the set of arcs excluded in subproblem k;

I_j = the set of arcs included in subproblem k;

The SVMRP Algorithm is as follows:

Step 1. [Initialization]. Initialize N and cost matrix C.
If city d is the depot, then O(d) = 0.
If city j is the origin of patient i and city k his destination, then O(j) = +i and O(k) = -i.
Exclude arcs that are infeasible by definition:
(-i,+i) = M
(0,-i) = M
(+i,0) = M.

Step 2. Form the tour T = ((0,+1),(+1,+2),...,
(+B-1,+B),(-1,-2),...,(-(B-1),-B),(-B,0))
Set CLUB* = Z_0.
Set BFSD* = T_0.

Step 3. Solve the modified assignment problem AS(1).
Test AS(1):
a. If node i is the immediate predecessor of the depot and O(i) < 0, then CHECK(|i|) = 1.
Set j = 2.
b. If S(1,2) = 0 for any i, AS(1) contains one or more subtours.
c. If O(i) < 0 and O(i) = -O(S(i,2)), then AS(1) contains an infeasible chain.
Repeat steps 3(b) and 3(c) for j=3,4,...,N-2.
If no subtours or infeasible chains are found, set CLUB* = Z_1, BFSD* = AS(1), and go to Step 9.
Set CLLB* = Z_1, E_1 = I_1 = 0, and k = P=1.

Step 4. [Branching setup]. If k > 1 and the path from the root node to subproblem k is $((1,i_1),(i_1,i_2),\ldots,(i_q,k))$, then for $j=i_1,i_2,\ldots,i_q$, define

m_j = the size of the smallest infeasibility in subproblem j,
u_j = the index v when subproblem j was created;
$v_j = \{r_1, r_2, \dots, r_{m_j}\}$, the set of the smallest infeasibility in subproblem j;
$L_j = \{(r_1,r_2),(r_2,r_3),\dots,(r_{m_j},r_1)\}$, the set of arcs in the smallest infeasibility.

Form $I_{i_1} = I_1 \cup \{(r_w, r_{w+1}) \mid w=1,\dots,u_{j-1}\}$
and $E_{i_1} = E_1 \cup \{(r_{u_j}, r_{u_{j+1}})\}$;
if $u_j = u_{j+1}$, $u_{j+1} = 1$.

Recursively form I_k and E_k.
Set v=1.

Step 5. [Branching]. Form $E_{P+v} = E_k \cup \{(r_v, r_{v+1})\}$ and, if $v > 1$, $I_{P+v} = I_k \cup \{(r_j, r_{j+1}) \mid j=1,\dots,v-1\}$.
Include the arcs $a_i \in I_{P+v}$ and exclude the arcs $a_i \in E_{P+v}$ in C.
If $I_{P+v} \cap E_{P+v} \neq \emptyset$, P+v is bounded; go to Step 7.
Solve the resulting MAP.
If the dual objective exceeds CLUB, P+v is bounded; exit the MAP subroutine and go to Step 7.

Step 6. [Bounding]. Determine the solution case:
a. If AS(P+v) contains a subtour or infeasible chain, save the shortest infeasibility, and insert P+v in the active subproblem list; if $Z_{P+v} <$ CLLB, reset CLLB = Z_{P+v} .
Go to Step 7.
b. If AS(P+v) is a tour, and $Z_{P+v} <$ CLUB, set CLUB = Z_{P+v} and BFSD = AS(P+v).

If CLUB ≤ CLLB*, the optimal has been found; go to Step 9.
Otherwise, go to Step 7.

Step 7. Set v = V+1.
If $v \le m_k$, go to Step 5.

Step 8. If Q = ∅, go to Step 9.
Otherwise, remove the first subproblem in Q, set k = P for that subproblem, and go to Step 4.

Step 9. Stop: BFSD* is the optimal solution, with objective value CLUB*.

4.3.3 A Computational Example. To illustrate the use of the SVMRP algorithm in solving the single-aircraft aeromedical routing problem, consider the problem depicted in Figure 4.34, where the symbol +n designates a patient origin, −n his corresponding destination, and 0 (Scott AFB) the depot. Great circle distances are given in Table 4.3.

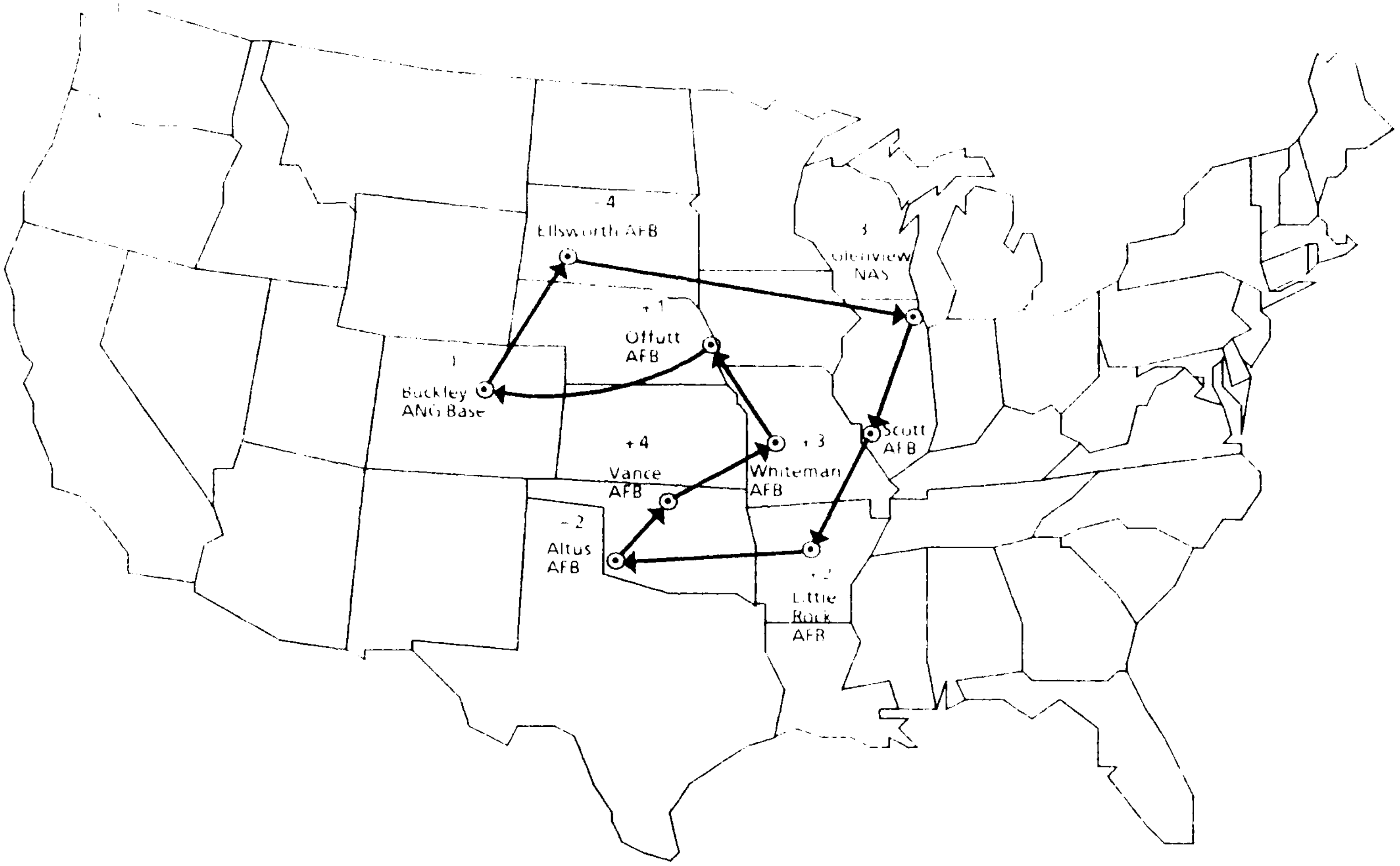
TABLE 4.3

DISTANCES[a] BETWEEN PATIENT ORIGINS AND DESTINATIONS

From Patient Service Point	To Patient Service Point								
	−3	0	+2	+3	+1	−4	−1	+4	−2
Glenview NAS	M	184	292	449	670	516	598	618	881
Scott AFB	184	M	195	310	540	357	514	434	697
Little Rock AFB	292	195	M	215	380	232	434	493	719
Whiteman AFB	449	310	215	M	288	200	566	787	790
Offutt AFB	670	540	380	288	M	211	436	814	632
Ellsworth AFB	516	357	232	200	211	M	381	642	697
Buckley ANGB	598	514	434	566	436	381	M	295	224
Vance AFB	618	434	493	787	814	642	295	M	320
Altus AFB	881	697	719	790	632	697	224	320	M

[a]Distances are in nautical miles, modulo 5.

Source: Global Navigation and Planning Chart, GNC-2N, Defense Mapping Agency, St Louis, MO., 1968.



OPTIMAL: 2595 NM

Figure 4.34. Patient Locations, Problem I

In our first trial, we used (symmetric) distances in nautical miles, a simple sequential upper bound

(0->+1->+2->+3->+4->-1->-2->-3->-4),

assignment relaxation lower bounding, and Bellmore-Malone-Murty subtour and infeasible chain elimination. The optimal objective for the same problem without order restrictions (the simpler Traveling Salesman Problem) is 2130 nautical miles, and the corresponding route is:

0->+3->+2->+4->-2->-1->-4->+1->-3->0.

The TSP tour is infeasible because the stop at Altus (-1) occurs before patient 1 is picked up at Offutt (+1). With Christofides-Balas lower bounding applied to the initial MAP, the TSP optimal was found at the root node.

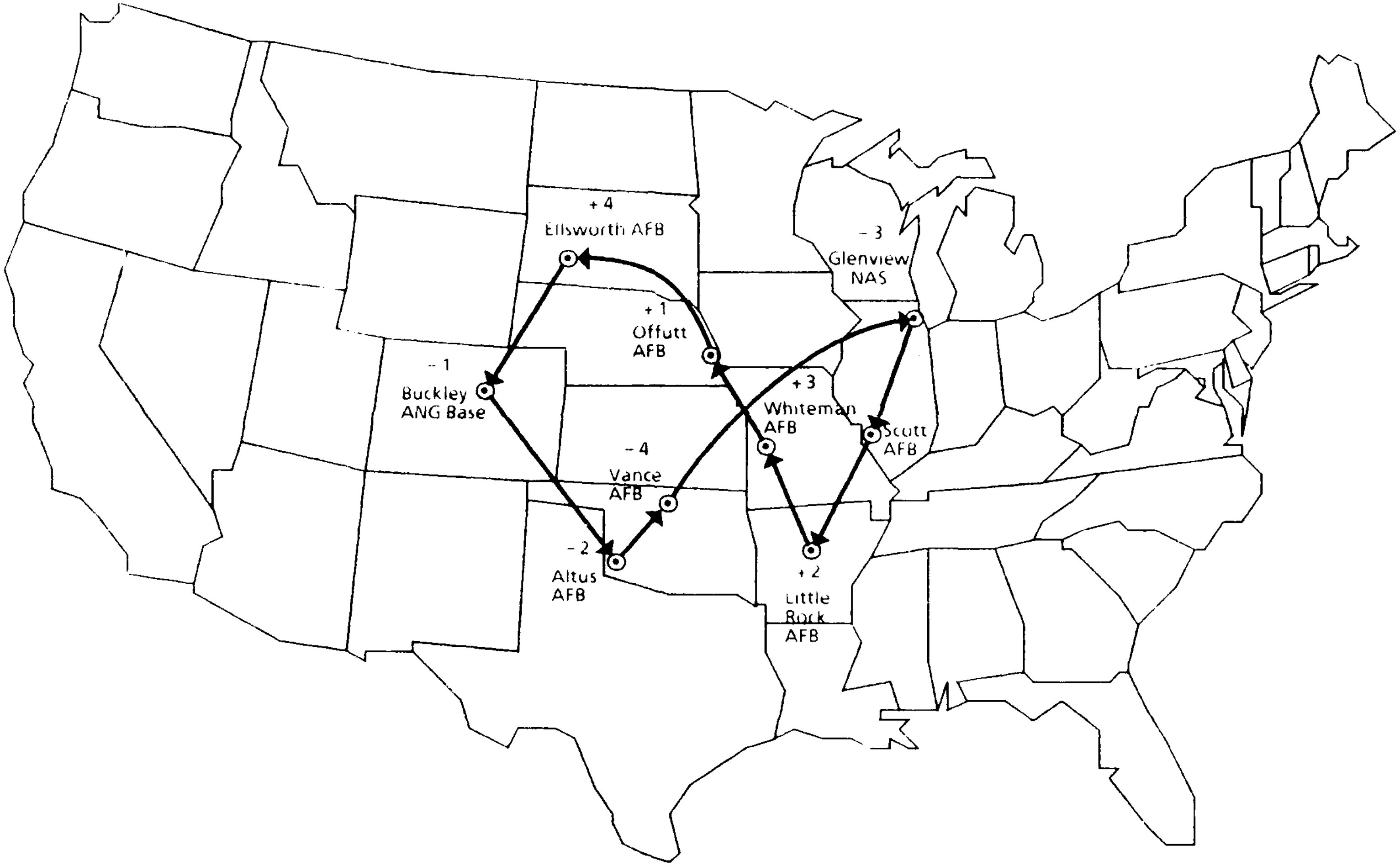
For the first trial problem, the SVRP algorithm generated initial upper and lower bounds of 3975 and 2015 miles respectively. The latter is the initial MAP solution value without Christofides-Balas lower bounding applied. The SVMRP optimal objective is 2595 nautical miles, and the route is:

0->+2->-2->+4->+3->+1->-1->-4->-3->0.

25 subproblems were created, including 2 feasible tours, 7 solutions containing subtours, 6 containing infeasible chain solutions, and 11 problems that either exited the MAP solution procedure when the value of the dual objective exceeded the incumbent upper bound or were skipped because of an arc inclusion-exclusion conflict.

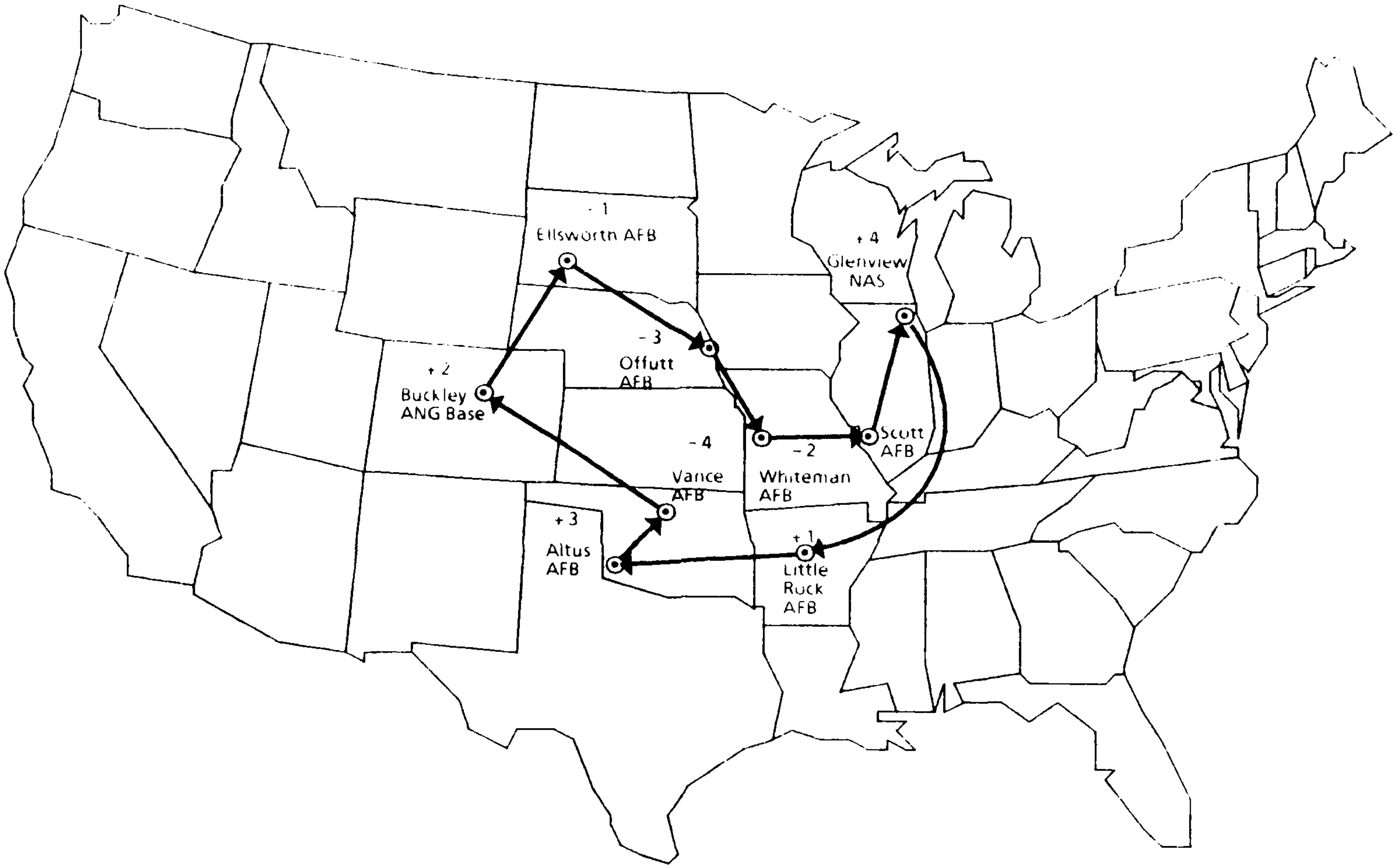
Two other trials depicted in Figures 4.35 and 4.36, show the effect of different origin-destination pairings. The new labels are given in Table 4.4. The results are summarized in Table 4.5. While it is difficult to explain performance with just a few solutions, these results are interesting in several respects. First, using a nearest neighbor upper bounding technique which allowed only feasibly ordered selections yielded a much better initial upper bound than the method of making all pickups first, then all deliveries. Secondly, relatively few subproblems were solved, with a significant proportion either fathomed because of an exclusion/inclusion conflict, terminated early in the MAP procedure for excessive dual objec-tive, or eventually fathomed by the incumbent, the first of which was found after solving relatively few subproblems. Third, the solution to problem II illustrates a fundamental difference between TSP and SVMRP, namely, that SVMRP optimal route segments can intersect, which will not occur in a Euclidian TSP.

The SVMRP algorithm is designed to handle non-unique origin-destinations pairs for different patients, as well as cases where the depot is either an origin, a destination, or both. The latter case is actually simple to handle, since that service requirement cannot be inmproperly sequenced. Both situations occur frequently in actual route planning.



OPTIMAL: 2595 NM

Figure 4.35. Patient Locations, Problem II.



OPTIMAL: 2425 NM

Figure 4.36. Patient Locations, Problem III.

TABLE 4.4

PATIENT ORIGIN/DESTINATION LABELS

Airfield	Node Labels		
	Problem		
	I	II	III
Glenview Naval Air Station	-3	-3	+4
Scott Air Force Base	0	0	0
Little Rock Air Force Base	+2	+2	+1
Whiteman Air Force Base	+3	+3	-2
Offutt Air Force Base	+1	+1	-3
Ellsworth Air Force Base	-4	+4	-1
Buckley Air National Guard Base	-1	-1	+2
Vance Air Force Base	+4	-4	-4
Altus Air Force Base	-2	-2	+3

TABLE 4.5

SVMRP COMPUTATIONAL RESULTS

	Problem		
	I	II	III
Bounding data			
Initial upper bound	3975*	2815	2690
Initial lower bound	2015**	2130	2130
Optimal objective	2595	2595	2425
Initial lower gap	580	465	295
Initial upper gap	1380	220	265
Solution performance data			
First subproblem with tour	11	16	18
Subproblems generated:			
Branched by the			
LIFO rule	3	1	1
Best bound rule	8	13	11
Exclusion conflict	5	7	3
Exceeded upper bound	6	6	11
during MAP solution	2	2	1
Incumbent optimal	1	5	4
Fathomed by an incumbent	1	5	4
Total	25	34	31

*Sequential upper bound rule used.

**Christofides-Balas lower bounding not used.

4.4 Summary and Conclusions. In this chapter, we have shown that the single-aircraft aeromedical routing problem can be solved with a branch and bound algorithm that employs a new technique for eliminating infeasible paths that violate the many-to-many ordering restriction that origins be visited before destinations. The method does not have the size limit or unique origin-destination requirement of the Psaraftis dynamic programming approach. Because the single aircraft model is only one stage of the complete aeromedical model, we have not attempted to generate optimal computer code, nor have we run exhaustive tests, particularly since no comparative solution data exists. Prof. Bodin used our data to run Sexton's algorithm without time windows and obtained the same optimal solution in an informal test. We strongly recommend further computational research. Also, we have not considered other problem characteristics, such as passenger priority and constraints on vehicle capacity and route length, which are important considerations in actual route planning and therefore should also be investigated.

Overall, we can conclude that solutions to single vehicle, many-to-many routing problems of significant size can be solved, with reasonable processing times. In fact, the number of subproblems required to solve the aeromedical problem is less, in the tests we have run to date, than the same problems without precedence relationships.

ENDNOTES

1. Psaraftis [PSAR78] uses a weighted composite of waiting and travel time to represent customer disutility, but no multiple objective models have appeared in the literature.

2. We did not ignore some particularly powerful heuristics and Lagrangean techniques. Indeed, techniques described by Golden et al [GOLD80] provide initial solutions, improve bounds, and find near-optimal solutions. As a matter of emphasis, we are interested primarily in exact methods.

3. Murty [MURT76] solves the MAP, rather than simply reducing the cost matrix, which (i), gives higher lower bounds at the expense of increased computation, and (ii), generates a complete, though possibly infeasible, solution. The Little algorithm will not provide a solution until at least n subproblems have been created, and then only if the procedure creates n successive arc inclusion branchings from the root node.

4. If $x_{12} = 1$, then if x_{21} were allowed to equal 1, a subtour would result. Therefore, the additional arc cut, $x_{21} = 0$, is implied to preserve feasibility. In general, each arc inclusion should also exclude the arc that would make the path containing the included arc a circuit.

5. The shortest subtour is the one with minimum cardinality. If several subtours have the same cardinality, arbitrarily select one. This rule of thumb produces the

fewest subproblem branches, but there is no theoretical evidence that it produces the fewest total subproblems for the entire solution process.

6. Admittedly, Gillete's code can be significantly improved. First, instead of exhaustively searching for the next unfathomed subproblem to solve, maintain a stack referencing unfathomed subproblems in ascending solution objective value order. Second, at a small expense in solution time, a fast heuristic can provide an initial upper bound close to the optimal (see [GOLD80]). And third, published AP algorithms are considerably faster than Gillette's version of the Hungarian algorithm. [CARP80a] Using Murty's [MURT76] labeling method, we eliminated a major theoretical mistake in Gillette's algorithm that caused it to cycle infinitely with certain data, and greatly improved its efficiency.

7. To include an arc (i,j), all $c_{.j}$ in row i and all $c_{i.}$ in row j are set to infinity except c_{ij}. Alternatively, solve the MAP with row i and column j deleted, and set $x_{ij} = 1$.

8. The same x_{ij}'s will be equal to 1, $c_{pq}' > c_{pq}$ if (p,q) is the arc to be prohibited, and the objective will be equal to Z_{NEW}.

9. For arc (i,j), arcs (a,i) and (j,b) link with arc (i,j). h_j, then, can be 0, 1 or 2.

10. Their method is not restricted to asymmetric problems, but they recommend the use of a matching algorithm to replace the assignment procedure for the symmetric case.

11. Earlier, we discussed the distinction between sequencing (or permutation) and scheduling problems. The latter type explicitly considers time in addition to arc weights. Restrictions such as desired pickup or delivery times create scheduling problems. Treatments of the Dial-a-Ride problem by Psaraftis [PSAR78] and Sexton [SEXT79] explicitly consider time. In the following, we will restrict our attention to sequencing problems exclusively.

12. In other words, each origin has a unique destination associated with it, and vice versa. We could allow several passengers to have exactly the same origin and destination, if we do not have capacity constraints.

13. Christofides [CHRI75] refers to this as the *open routing problem*, as opposed to the *closed problem* of finding tours.

14. To *insert* node c between nodes a and b, replace the arc (a,b) with ((a,b),(c,b)), in which a and c *precede* b.

15. Open tours do not return to the depot after the last patient is delivered.

CHAPTER V

THE MULTIPLE VEHICLE, MULTIPLE DEPOT, MANY-TO-MANY ROUTING PROBLEM (MVMDMRP)

5.1 Introduction. In Chapter IV, we completed the first two versions of the aeromed model. In this chapter, we expand Version II of the aeromedical model to handle multiple aircraft and then complete Version III development by incorporating multiple depots. The last two sections present a computational example of the multiple depot algorithm, and present algorithm extensions that observe origin-destination precedence relationships.

5.2 The Multiple Vehicle Routing Problem. The multiple vehicle routing problem (MVRP) is a straightforward generalization of the single vehicle problem. Gavish and Srikanth define the simplest MVRP, the Multiple Traveling Salesman Problem (MTSP), as follows:

> "Given a set of n cities, find a set of routes for m salesman starting from and ending at the base city 1, such that each city (apart from city 1) is visited by one and only one salesman." [GAVI80]

MTSP variations assume constant or variable values of m, and may require the matrix $C = [c_{ij}]$ to be either symmetric or asymmetric, or allow it to be either.

5.2.1 MVRP Problem Statement. Because the aeromedical problem involves a fixed number of vehicles and asymmetric costs (flying times affected by winds), we assume that m is constant and C = $[c_{ij}]$ can be either asymmetric[1] or symmetric in the following formulation:

Problem MTSP:

Find values of the variables X = $[x_{ij}]$ that

$$\text{Minimize} \quad \sum_{i=1}^{r} \sum_{j=1}^{r} c_{ij} x_{ij} \qquad r = m + n; \tag{5.1}$$

$$\text{subject to} \quad \sum_{i=1}^{r} x_{ij} = 1, \qquad j = 1, 2, \ldots, r; \tag{5.2}$$

$$\sum_{j=1}^{r} x_{ij} = 1, \qquad i = 1, 2, \ldots, r; \tag{5.3}$$

$$X = [x_{ij}] \in T \tag{5.4}$$

$$X_{ij} = 0 \text{ or } 1 \tag{5.5}$$

Golden et al [GOLD77] note that

> Three different papers ([BELL71],[ORLO74],[RUSS77]) published in 1973 and 1974 independently derived equivalent TSP formulations of the MTSP and consequently showed that the M-salesman problem is no more difficult than its one-salesman counterpart.

To obtain this equivalence, the c_{ij} in (5.1) are defined as shown in Figure 5.1 for the MTSP. The first m-1 rows and columns are copies of row 1 and column 1 respectively. This creates, in effect, m-1 duplicates of the original depot. The (m-1)x(m-1) submatrix of infinite costs prohibits arcs between duplicates of the depot, which forces all m vehicles to be used.[2] All other costs remain the same.

$$\begin{array}{c|c|c}
 & 1\ 2\ .\ .\ .\ m-1 & m\ \ m+1\ .\ .\ .\ .\ .\ .\ .\ .\ .\ .\ m+n-1 \\
\hline
\begin{array}{c}1\\2\\ .\\ .\\ .\\ m-1\\ m\end{array} & C'_{ij} = \infty & C'_{ij} = C_{ij} \\
\hline
\begin{array}{c}m+1\\ .\\ .\\ .\\ m+n-1\end{array} & C'_{ij} = C_{ij} & C'_{ij} = C_{ij}
\end{array}$$

Figure 5.1. MTSP cost matrix.

Constraints that prohibit subtours are similar for TSP and MTSP. Let nodes 1,2,...,m denote the m duplicates of the depot, I_0 = {1,2,...,m} be the set of depot indices, and N={1,2,...,m,m+1,...,r} be the set of indices for all nodes. MTSP subtour elimination constraints are [GOLD77]

$$T = \{[X_{ij}]: \sum_{i\varepsilon Q} \sum_{j\varepsilon Q} X_{ij} \leq 1 \quad \textit{for every } Q \subset N - I_0\}; \tag{5.4a}$$

$$T = \{[X_{ij}]: \sum_{i\varepsilon Q} \sum_{j\varepsilon Q} X_{ij} \leq |Q| \quad \textit{for every } 0 \neq Q \subset N - I_0\}; \tag{5.4b}$$

$$T = \{[X_{ij}] : Y_i - Y_j + (n-m)X_{ij} \leq n-m-1, \ i \neq j, \ i,j \varepsilon N - I_0\} \tag{5.4c}$$

5.2.2 <u>MVRP Solution Methods</u>. Table 5.1 provides information on MTSP research reported in the literature. Our solution method most closely resembles that of Svestka and Huckfeldt

TABLE 5.1

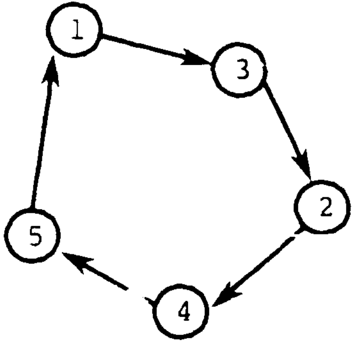
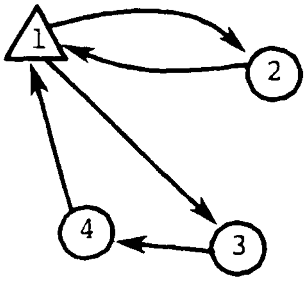
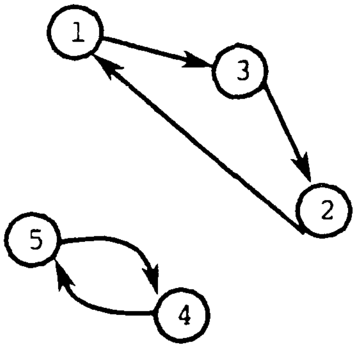
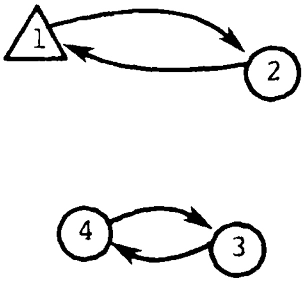
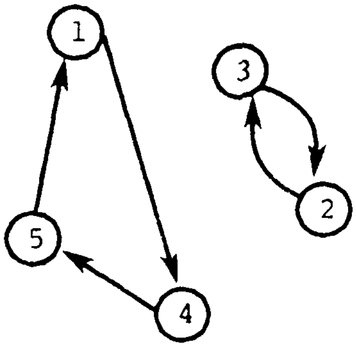
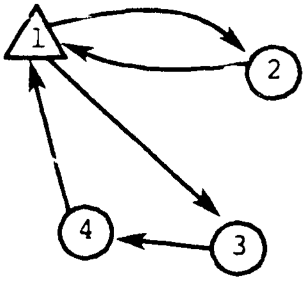
MULTIPLE TSP SOLUTION METHODS

	Restrictions on		
Researchers	c_{ij}	m	Solution Approach
Ali and Kennington [ALI 79]	Symmetric	Variable	Lagrangean relaxation, extending the Held and Karp l-tree method to multiple l-trees (m-trees).
Bellmore andHong [BELL71]	Symmetric	Constant	No method specified.
Svestka and Huckfeldt [SVES73]	None	Constant	TSP transformation using Bellmore-Malone branching, AP relaxation lower bounding, and heuristic upper bounding.
Gavish and Srikanth [GAVI80]	Symmetric	Constant	Branch and bound, employing Lagrangean relaxation, minimal spanning m-tree, and subgradient optimization techniques.
Laporte and Norbert [LAPO80]	None	Variable	LP relaxation.
Hong and Padberg [HONG77]	Symmetric	Variable	Not reported.
Russell [RUSS77]	Symmetric	Variable	Heuristic, based on Gillette's TSP sweep procedure; additional restrictions can be included.

[SVES73], which uses a Bellmore-Malone TSP branch-and-bound procedure with three modifications. First they obtain an initial upper bound heuristically from the root assignment relaxation solution. Secondly, to branch a subproblem, they use a modified primal-dual transportation algorithm to exploit the fact that the solution of any parent problem is dual feasible for all of its subproblems.[3]

The third modification interprets the assignment solution obtained from the MTSP cost matrix. Suppose n=4, m=2, and r=5. Three solution cases are possible (Figure 5.2). Cases I and III are (equivalent) feasible solutions that would not require further branching, while Case II is an infeasible subtour solution requiring that the shortest (disconnected) subtour not containing a copy of the depot be branched. In general, all MTSP tours are feasible (Case I), while all subtour solutions are either feasible (Case II) or infeasible (Case III), depending on whether or not every subtour contains a depot node.

Svestka and Huckfeldt note in their article that the presence of multiple vehicles does not increase solution time; in fact, for problems of a given size, the MTSP seems to converge to optimality more quickly. [SVES73] Intuitively, it appears that the feasibility of Case III solutions increases the likelihood that a given solution will not need to be branched. However, in the absence of any theoretical investigation, the relative efficiency of TSP

Case	MARP Solution	Interpretation
I	 Tour	 Feasible Tours
II	 Subtours	 Infeasible Subtours
III	 Subtours	 Feasible Tours

△ Depot

Source: [SVES73]

Figure 5.2 Solution cases for the Multiple Traveling Salesman Problem

vs. MTSP algorithms for solving the same problems remains an open question.

The Christofides-Balas bounding procedures discussed in the last chapter can be extended easily to the MTSP case, provided cutsets, coverings, and articulation points are properly chosen. Only subtours that do not contain a depot should be used. In Case II, Figure 5.2, for example, one subtour, ((1,3),(3,2),(2,1)), is feasible, while the other is not. Only nodes 4 and 5 should be used to find reachable node sets to construct cutsets, and we should only search for an articulation point for subtour ((4,5),(5,4)).

5.2.3 An MVRP Computational Example. For this example, we changed the Gillette ten-city cost matrix from the last chapter to make it correspond to Figure 5.1. The only significant modification to the TSP algorithm necessary to solve Problem MTSP is to restrict the branching routine to select the shortest subtour that does not visit any copy of the depot. In each iterative step, a solution is produced that corresponds to one of five cases:

Case	Solution Type
I	Feasible (Single tour)
II	Infeasible (Disconnected subtour(s))
III	Feasible (All subtours connected)
IV	Incomplete (dual objective exceeded upper bound)
V	Infinite (Arc inclusion/exclusion conflict)

The two additional cases (IV and V) result from improvements to the Bellmore-Malone algorithm discussed in the last chapter.

For three vehicles, we obtained the solution depicted in Figure 5.3. Solution details are provided in Table 5.2 Fewer subproblems were solved (27 vs. 119) than for Problem TSP, but the total distance of the optimal solution is more than thirty per cent greater (3749 vs. 2855). Of those 27 solutions, the first 14 contained subtours. However, once the algorithm found an incumbent optimal in the 15th subproblem, subsequent subproblems were either optimal (2 cases), terminated in the MAP procedure because the dual objective exceeded the upper bound (9 cases), or excluded because of an arc exclusion conflict (2 cases).

5.3 The Multi-Depot Routing Problem (MDRP). In the MVRP model, we assumed m vehicles were available at a single depot. In that model the problem is essentially one of creating a partition of the n patients and then constructing r routes. The multiple depot problem is similar, in that each vehicle assigned to one of k depots must be routed to service a subset of the n patients. In this section, we will formally state the multi-depot problem, discuss approaches reported in the literature, and then formulate a new algorithm for solving the problem.

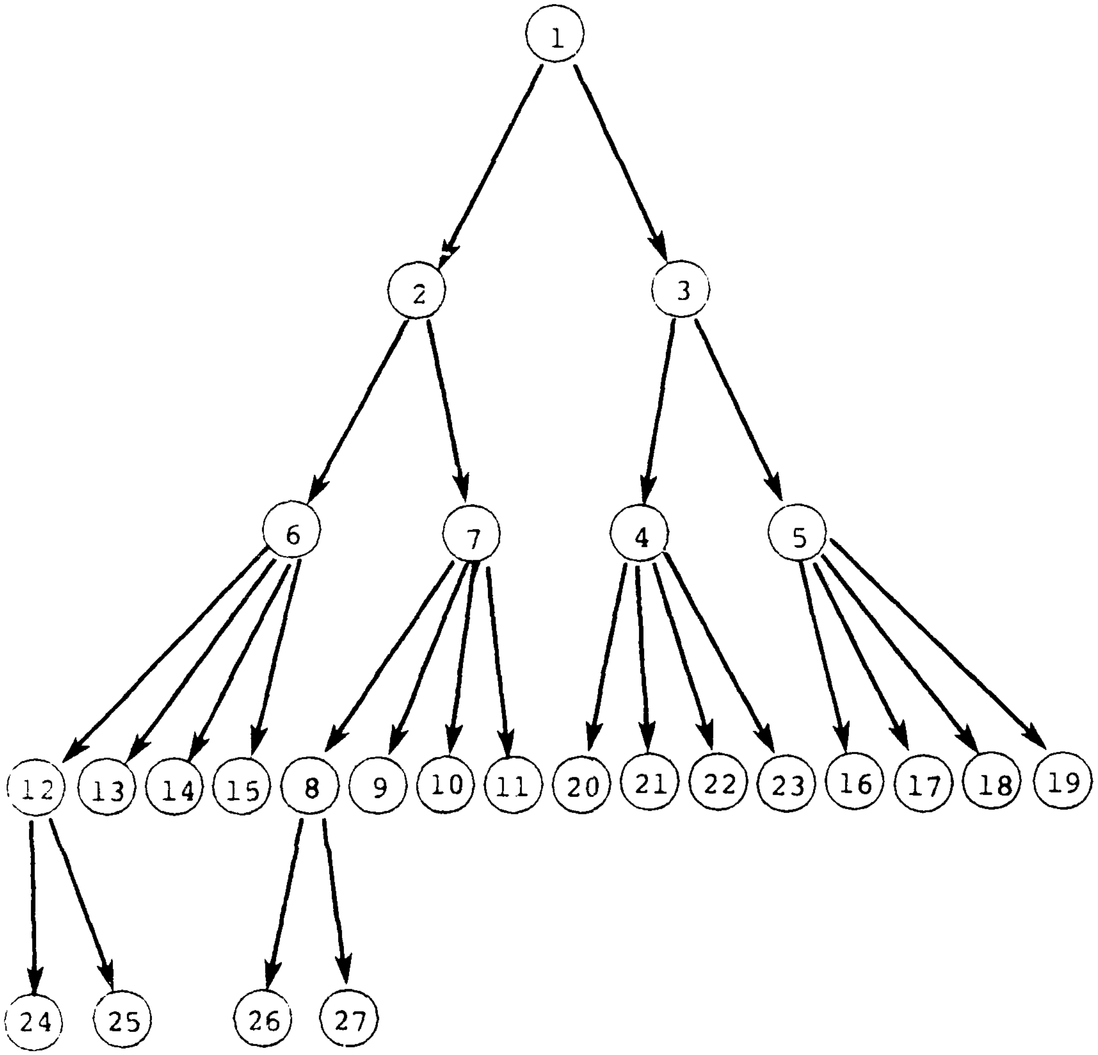


Figure 5.3. Solution tree for a 10-city, 3-vehicle problem.

5.3.1 MDRP Problem Formulation. For purposes of discussion, we assume that:

1) One aircraft is assigned to each of the m depots.

2) Each aircraft must be used, and it must start and end its route at its assigned depot.

3) A route cannot visit more than one depot.

In the MDRP formulation, we still interpret r to be the total number of stops, but we now define m as the number of

TABLE 5.2

SOLUTION DATA FOR A 10-CITY, 3-VEHICLE PROBLEM

Subproblem	Optimal Solution	Arcs		
		Excluded	Included	Solution[a]
1	3402	-	-	II: 12->10->12
2	3450	(12,10)	-	II: 8->7->8
3	3450	(10,12)	(12,10)	II: 8->7->8
4	3594	(8,7),(10,12)	(12,10)	II: 11->12->10->9->11
5	3594	(7,8),(10,12)	(8,7),(12,10)	II: 11->12->10->9->11
6	3594	(8,7),(12,10)	-	II: 10->12->11->9->10
7	3594	(7,8),(12,10)	(8,7)	II: 10->12->11->9->10
8	3642	(7,8),(10,12), (12,10)	(8,7)	II: 12->11->12
9	3842	(7,8),(12,10), (12,11)	(8,7),(10,12)	II: 11->10->12->9->11
10	3825	(7,8),(11,9), (12,10)	(8,7),(10,12), (12,11)	II: 12->11->10->12
11	3750	(7,8),(9,10), (12,10)	(8,7),(10,12), (11,9),(12,11)	II: 8->7->6->8
12	3642	(8,7),(10,12), (12,10)	-	II: 12->11->12
13	3842	(8,7),(12,10), (12,11)	(10,12)	II: 11->10->12->9->11
14	3825	(8,7),(11,9), (12,10)	(10,12),(12,11)	II: 12->11->10->12
15	3749	(8,7),(9,10), (12,10)	(10,12),(11,9), (12,11)	I[b]
16	≥3842	(7,8),(10,12), (11,12)	(8,7),(12,10)	IV
17	9999	(7,8),(10,12), (12,10)	(8,7),(11,12), (12,10)	V

TABLE 5.2 (Cont.)

SOLUTION DATA FOR A 10-CITY, 3-VEHICLE PROBLEM

Subproblem	Optimal Solution	Arcs		
		Excluded	Included	Solution[a]
18	3749	(7,8),(10,9), (10,12)	(8,7),(11,12), (12,10)	I[c]
19	≥3874	(7,8),(9,11), (10,12)	(8,7),(10,9), (11,12),(12,10)	IV
20	≥3842	(8,7),(10,12), (11,12)	(12,10)	IV
21	9999	(8,7),(10,12), (12,10)	(11,12),(12,10)	V
22	≥3750	(8,7),(10,9) (10,12)	(11,12),(12,10)	IV
23	≥3835	(8,7),(9,11) (10,12)	(10,9),(11,12), (12,10)	IV
24	≥3890	(8,7),(10,12), (12,10),(12,11)	-	IV
25	≥3788	(8,7),(10,12), (11,12),(12,10)	(12,11)	IV
26	≥3890	(7,8),(10,12), (12,10),12,11)	(8,7)	IV
27	≥3788	(7,8),(10,12), (11,12),(12,10)	(8,7),(12,11)	IV

[a] Solution data includes the case, and the uninterpreted subtour, if applicable.

[b] The interpreted incumbent solution is: Vehicle 1: 1->8->10->9->7->5->6->4->1 Vehicle 2: 1->3->1 Vehicle 3: 1->2->1

[c] The interpreted incumbent solution is: Vehicle 1: 1->3->1 Vehicle 2: 1->4->6->5->7->9->10->8->1 Vehicle 3: 1->2->1

depots and n the number of patient origins and destinations. Because we do not create duplicate copies of any depot, we do not have to interpret the solution as we did for Problem MTSP. However, if we relax the first assumption, we would have to include Problem MTSP depot copying and solution interpretation procedures.

Problem MDRP:

$$\textit{minimize } Z = \sum_{i=1}^{r} \sum_{j=1}^{r} C_{ij} X_{ij} \qquad (r = m+n); \tag{5.6}$$

$$\textit{subject to:} \quad \sum_{i=1}^{r} X_{ij} = 1. \qquad j = 1,2,\ldots,r; \tag{5.7}$$

$$\sum_{j=1}^{r} X_{ij} = 1, \qquad i = 1,2,\ldots,r; \tag{5.8}$$

$$X = [X_{ij}] \varepsilon T \tag{5.9}$$

$$X_{ij} = 0 \textit{ or } 1 \tag{5.10}$$

Constraints (5.9) are similar to those of Problem MTSP. Let nodes 1,2,...,m denote the m different depots, I_0={1,2,...,m} be the set of depot indices, and N = {1,2,...,m,m+1,...,r} be the set of all node indices, where r = m + n. and destinations. Then, let

$$T = \{[X_{ij}]: \sum_{i \varepsilon Q} \sum_{j \notin Q} X_{ij} \geq 1 \textit{ for every } Q \subset N - I_0\}; \tag{5.9a}$$

$$T = \{[X_{ij}]: \sum_{i \varepsilon Q} \sum_{j \varepsilon Q} X_{ij} \leq |Q| - 1 \textit{ for every } 0 \neq Q \subset N - I_0\}; \tag{5.9b}$$

$$T = \{[X_{ij}] \mid Y_i - Y_j + (n-m)X_{ij} \leq n - m - 1,\ i \neq j,\ i,j \varepsilon N - I_0\}. \tag{5.9c}$$

Subtour elimination constraints (5.9) eliminate two types of infeasibility. [GOLD77] Fi~ure 5.4 shows the three MDRP solution cases. Only Case III is feasible; the solution

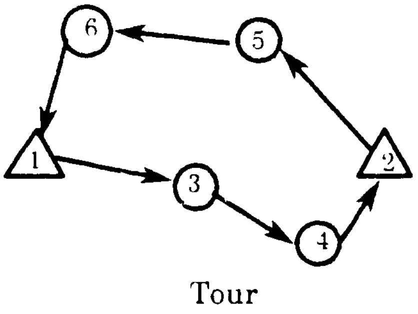
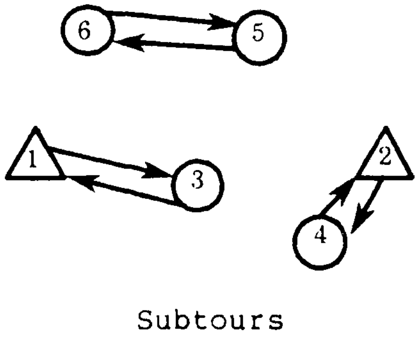
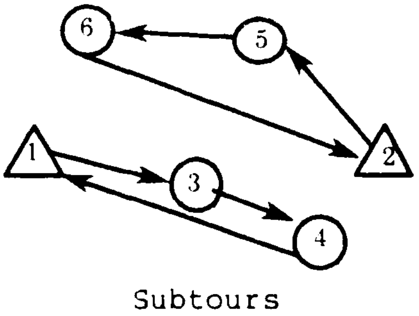
Case	MDRP Solution	Interpretation
I		Infeasible
II		Infeasible
III		Feasible

Figure 5.4 Solution Cases for the Mu]tiple Depot Routing Problem

consists of subtours that each contain one and only one depot. Tours (Case I) are infeasible because they contain more than one depot, which violates the stipulation that a given aircraft depart and return to the same depot. Case II represents an infeasible solution, because at least one (disconnected) subtour contains no depot nodes.

5.3.2 MDRP Solution Methods. Bodin et al [BODI83] discuss three heuristic approaches to the MDRP. The first, the cluster first-route second procedure, initially groups nodes together into clusters around each depot, and then finds routes for each cluster using either heuristic or optimal routing methods. Gillette and Johnson [GILL76b], for example, employ the heuristic Gillette-Miller sweep algorithm [GILL74] to first find clusters, and then create a set of routes through the clusters. Finding optimal clusters is an open problem. [BODI83] The second approach, called route first-cluster second, first establishes single a TSP tour through all nodes, then breaks up the single route into a set of routes that each include one depot.

The third, the Tillman-Cain savings algorithm [TILL72], based on the well-known Clarke-Wright heuristic [CLAR64], associates nodes with depots on the basis of savings. The concept of savings is as follows. If we assume that we have as many vehicles as depots, then a feasible solution serves each non-depot node via a round trip from the nearest depot. If we can find instances where a three-arc route requires less distance to serve two customers than the two round trips to serve them separately, we will save

$$S_{1,2} = d_{0,1} + d_{0,2} - d_{1,2} \tag{5.11}$$

where 0, 1 and 2 index the depot and the two customer nodes respectively, and $d_{i,j}$ is the distance from node i to node j. The Clarke-Wright procedure can also observe restrictions such as vehicle capacity, maximum number of vehicles available (with variable capacities), and maximum number of stops on one route.

For the multi-depot case, the savings feature must be modified. To illustrate, suppose we have the following two-city, two-depot problem.

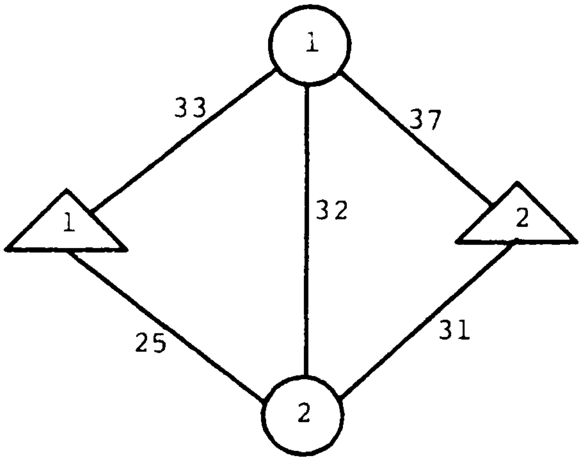


Figure 5.5. Two-Depot Problem.

If we define $S_{ij}{}^{D_k}$ as the savings realized by including stops at nodes i and j on a route out of depot D_k, then the Clarke-Wright savings would be

$$S^{1}_{1,2} = d_{1,1} + d_{1,2} - d_{1,2} = 33 + 25 - 32 = 26$$

$$S^{2}_{1,2} = d_{2,1} + d_{2,2} - d_{1,2} = 37 + 31 - 32 = 36$$

Servicing both points from depot 2 generates larger "savings", but the total distance is greater because the Clarke-Wright savings function fails to take into account which depot is closer to the nodes to be served.

To correct this, Tillman and Cain define

$$\overline{d_i^k} = min\{d_i^m\} - (d_i^k - min\{d_i^m\}) \tag{5.12}$$

as the 'modified' distance between depot k and point i. Then, by defining savings as

$$\overline{S_{ij}^k} = \overline{d_i^k} + \overline{d_j^k} - d_{ij} \tag{5.13}$$

the modified distances reduce savings by the amount that the actual distance from depot k to, say node i, exceeds the distance from the nearest depot to node i. Computing these revised distances and savings:

$$\overline{d_1^1} = 33 - (33 - 33) = 33$$

$$\overline{d_1^2} = 33 - (37 - 33) = 29$$

$$\overline{d_2^1} = 25 - (25 - 25) = 25$$

$$\overline{d_2^2} = 25 - (31 - 25) = 19$$

$$\overline{S^{1}_{1,2}} = \overline{d_1} + \overline{d_2^2} - d_{1,2} = 33 + 25 - 32 = 26$$

$$\overline{S^{2}_{1,2}} = \overline{d_1^2} + \overline{d_2^2} - d_{1,2} = 29 + 19 - 32 = 16$$

TABLE 5.3

TILLMAN-CAIN AND CLARK-WRIGHT SAVINGS

	Servicing Depot	
	D^1	D^2
Round trip distances	116	136
Clark-Wright savings	26	36
Single-trip distances	90	100
Tillman-Cain savings	26	16

Table 5.3 shows that the Tillman-Cain technique would save ten more units than the Clark-Wright solution.

In their article, Tillman and Cain execute their procedure as a heuristic. While they claim that an optimal solution could be obtained, we have not seen any results reported in the literature. Because such restrictions as the number and capacity of vehicles can be explicitly handled, their method is potentially useful for solving problems more complex than the simple MDRP, and as a heuristic upper bounding device in optimization algorithms.

We propose a new procedure for solving Problem MDRP using Bellmore-Malone methods to branch the shortest subtour that does not contain a depot (Case II, Figure 5.4), or a tour (Case I, Figure 5.4). Problem MDRP is obviously similar to Problem MTSP, but does not require depot copying, and hence, does not require a solution interpretation procedure.

5.3.3 An MDRP Computational Example. For this example, we use the following distance matrix:

	ASF_1	ASF_2	1	2	3	4	5	6
ASF_1	M	M	33	45	32	68	25	20
ASF_2	M	M	37	27	25	24	31	56
1	33	37	M	15	14	60	32	48
2	45	27	15	M	16	51	36	58
3	32	25	14	16	M	46	21	42
4	68	24	60	51	46	M	46	65
5	25	31	32	36	21	46	M	24
6	20	56	48	58	42	65	24	M

Figure 5.6. Six-city, two-depot problem.
Source: [TILL72]

Our branch and bound algorithm generated the solution depicted in Figure 5.7 and the results shown in Table 5.4. Twenty nine subproblems were solved; only nine required further branching. Of those nine branched, Christofides-Balas lower bounding did not provide significant improvement. Gaps between MAP optimal solutions and actual MDRP optimal were only reduced nineteen per cent on average.

5.3.4 MDRP with Multiple Vehicles (MDMVRP). Relaxing the MDRP assumption of a single aircraft at each depot to one that specifies a given number at each depot (not necessarily the same number) does not require extensive algorithm modification. Rather, by using the depot copying feature of the MTSP model, and by enforcing the MTSP feasibility criterion that a feasible subtour must depart and

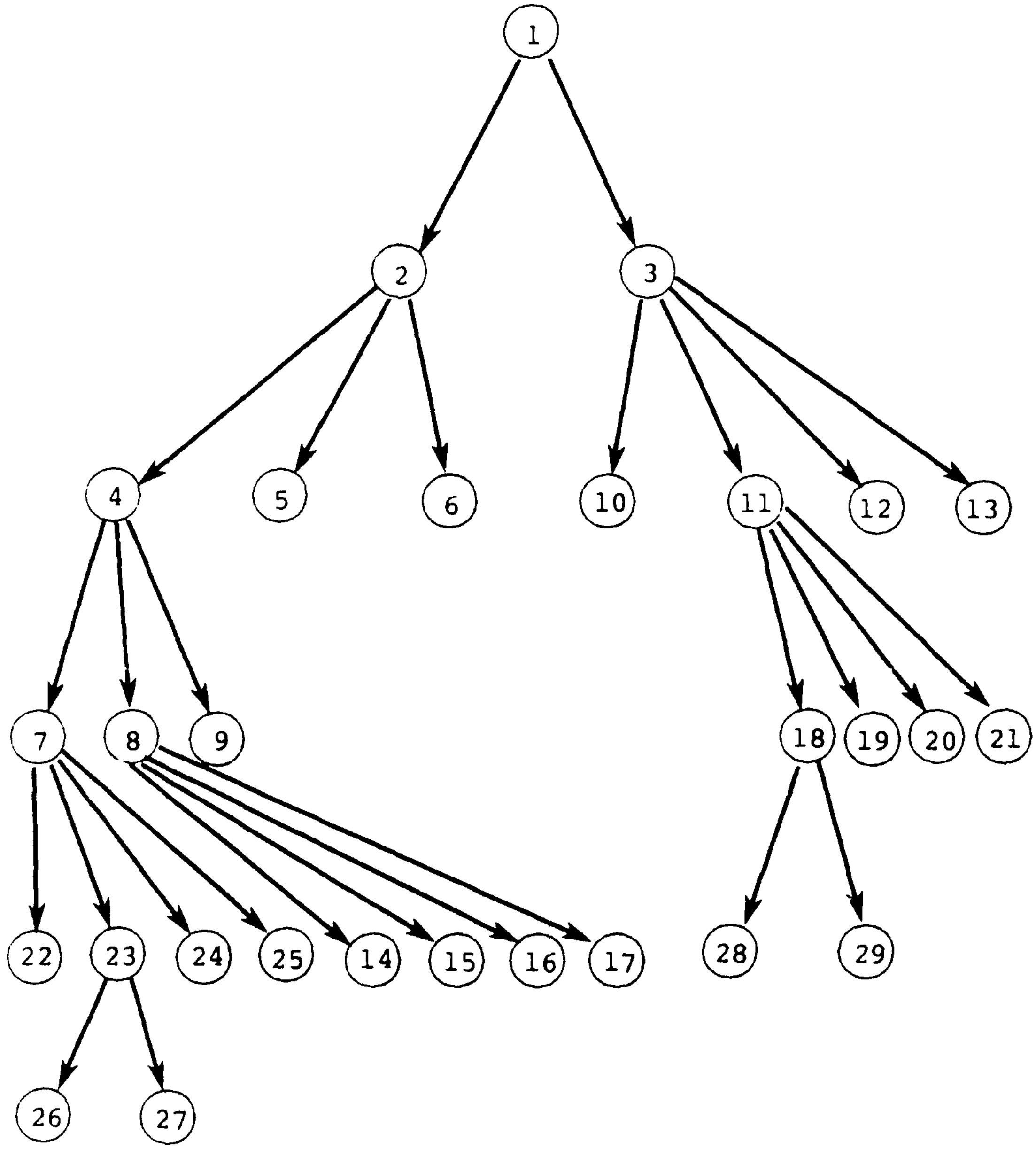


Figure 5.7. Solution tree for a 6-city, 2-depot problem.

return to a copy of the same depot, solving problem MDMVRP is straightforward. Only minor changes in bookkeeping steps and infeasibility testing are needed. If the number ot aircraft assigned to a depot is to be determined, then a new procedure must be devised.

TABLE 5.4

SOLUTION DATA FOR A 6-CITY, 2-DEPOT PROBLEM

Subproblem	Optimal Solution	Arcs		
		Excluded	Included	Solution[a]
1	165	-	-	II: 7->5->7
2	165	(7,5)	--	II: 5->4->3->5
3	172	(5,7)	(7,5)	II: 7->5->4->3->4->7
4	165	(5,4),(7,5)	-	II: 4->5->3->4
5	184	(4,3),((7,5)	(5,4)	II: 5->4->5
6	187	(3,5),(7,5)	(5,4),(4,3)	III[b]
7	174	(4,5),(5,4), (7,5)	-	II: 5->7->4->3->5
8	173	(5,3),(5,4), (7,5)	(4,5)	II: 7->5->4->3->7
9	186	(3,4),(5,4), (7,5)	(4,5),(5,3)	II: 7->4->5->3->7
10	9999	(7,5),(5,7)	(7,5)	V
11	174	(5,4),(5,7)	(7,5)	II: 4->7->5->3->4
12	≥206	(4,3),(5,7)	(7,5),(5,4)	IV
13	177	(3,7),(5,7),	(4,3),(7,5), (5,4)	III[c]
14	9999	(7,5),(5,4), (5,3),(5,7)	(4,5)	V
15	≥187	(5,7),(5,3), (5,4),(7,5)	(4,5)	IV
16	177	(7,3),(5,3), (5,4),(7,5)	(5,7),(4,5)	III[d]
17	≥197	(3,4),(5,3), (5,4),(7,5)	(5,7),(4,5), (7,3)	IV
18	176	(5,4),(4,7), (5,7)	(7,5)	II: 4->3->4

TABLE 5.4 (Cont.)

SOLUTION DATA FOR A 6-CITY, 2-DEPOT PROBLEM

Subproblem	Optimal Solution	Arcs		
		Excluded	Included	Solution[a]
19	9999	(7,5),(5,4),(5,7)	(4,7),(7,5)	V
20	≥205	(5,3),(5,4),(5,7)	(4,7),(7,5)	IV
21	≥199	(3,4),(5,4),(5,7)	(4,7),(7,5),(5,3)	IV
22	≥178	(4,5),(5,4),(5,7) (7,5)	-	IV
23	175	(4,5),(5,4),(7,4) (7,5)	(5,7)	II: 4->3->4
24	≥217	(4,3),(4,5),(5,4) (7,5)	(5,7),(7,4)	IV
25	≥198	(3,5),(4,5),(5,4) (7,5)	(4,3),(5,7),(7,4)	IV
26	≥201	(4,3),(4,5),(5,4) (5,7),(7,5)	(5,7)	IV
27	≥187	(3,4),(4,5),(5,4) (7,4),(7,5)	(4,3),(5,7)	IV
28	≥187	(4,3),(4,7),(5,4) (5,7)	(7,5)	IV
29	≥191	(3,4),(4,7),(5,4) (5,7)	(7,5),(4,3)	IV

[a] Solution data includes the case, and the uninterpreted subtour, if applicable.

[b,c,d] The interpreted incumbent solutions are:

	Vehicle 1:	Vehicle 2:
Subproblem 6:	ASF_1->3->2->1->5->6->ASF_1	ASF_2->4->ASF_2
Subproblem 13:	ASF_1->6->5->3->2->1->ASF_1	ASF_2->4->ASF_2
Subproblem 16:	ASF_1->1->2->3->5->6->ASF_1	ASF_2->4->ASF_2

5.3.5 MDRP with Precedence Relationships (MDMRP). With the introduction of precedence relationships, Problem MDRP must be changed to include order as a feasibility restriction. Since MDMRP is a multi-depot problem, only three solution cases are possible and only one case is feasible. Further, as Figure 5.8 illustrates, two additional infeasibilities can arise because of precedence.

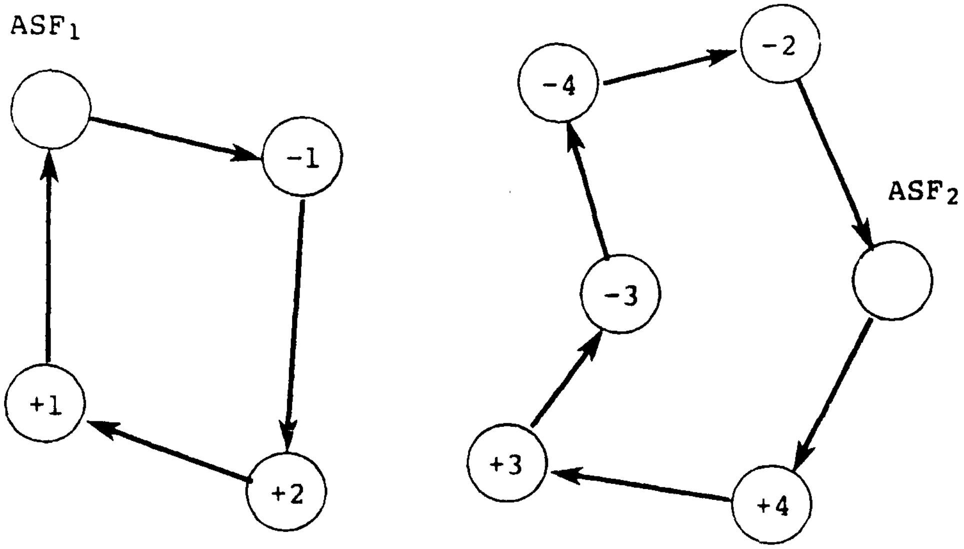


Figure 5.8. An MDMRP solution with precedence infeasibilities.

The first infeasibility arises when different aircraft serve the origin and destination of a patient, assuming no travel between depots.[4] In the example above, the aircraft serving ASF$_1$ visits patient 2's origin, while the second aircraft visits his destination. To eliminate this first infeasibility, we considered two branching strategies.

The first approach (Figure 5.9) is to solve two subproblems. One excludes the patient's origin node from the

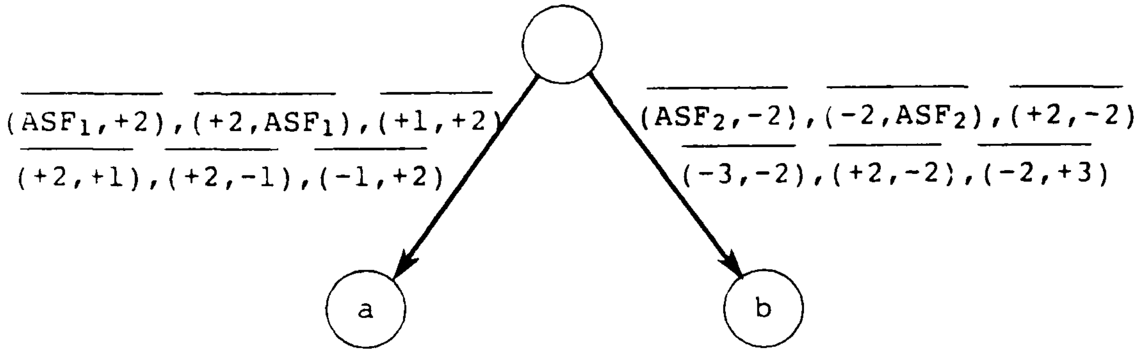

$\overline{(ASF_1,+2)},\overline{(+2,ASF_1)},\overline{(+1,+2)}$
$\overline{(+2,+1)},\overline{(+2,-1)},\overline{(-1,+2)}$
$\overline{(ASF_2,-2)},\overline{(-2,ASF_2)},\overline{(+2,-2)}$
$\overline{(-3,-2)},\overline{(+2,-2)},\overline{(-2,+3)}$
a
b

Figure 5.9. MDRP incomplete service exclusion branching.

subtour in which it occurs, and the other excludes his destination node. Arcs excluded on one branch are allowed on the other, and the arc exclusions clearly eliminate these two subtours from descendent subproblems. However, if we examine the pattern of exclusions closely, we can see that this branching does not partition the solution space. Given the following exclusion matrices, the solution

$$ASF_1 \text{ --> } +4 \text{ --> } +3 \text{ --> } -3 \text{ --> } -4 \text{ --> } ASF_1$$
$$ASF_2 \text{ --> } +2 \text{ --> } -2 \text{ --> } +1 \text{ --> } -1 \text{ --> } ASF_2$$

is in both subspaces, so this method does not partition.

To partition the solution subspace, we must ensure that a chain servicing a patient contains both his origin and destination nodes. (We consider the additional requirement of chain feasibility for origin-destination order below). The shortest of the two chains in Figure 5.11 can be eliminated by excluding arc (+2,-1) on the left branch and including it on the right, guaranteeing a partition by mutual exclusion. The two arc cuts prohibit the subtour

	ASF$_1$	ASF$_2$	+1	+2	+3	-1	-2	-3
ASF$_1$	∞	∞		M		∞	∞	∞
ASF$_2$	∞	∞				∞	∞	∞
+1	∞	∞	∞	M				
+2	∞	∞	M	∞		M		
+3	∞	∞			∞			
-1			∞	M		∞		
-2				∞			∞	
-3					∞			∞

(a) Arc exclusions for subproblem a.

	ASF$_1$	ASF$_2$	+1	+2	+3	-1	-2	-3
ASF$_1$	∞	∞		M		∞	∞	∞
ASF$_2$	∞	∞				∞	∞	∞
+1	∞	∞	∞	M				
+2	∞	∞	M	∞				
+3	∞	∞			∞		M	
-1			∞	M		∞		
-2		M		∞	M		∞	M
-3					∞		M	∞

(b) Arc exclusions for subproblem b.

Figure 5.10. Non-partitioning exclusions.

containing the chain in descendent subproblems. This second branching scheme will not force nodes +2 and -2 into the same subtour in descendent subproblems. Rather, it only guarantees that a particular incomplete service chain will be prohibited. The same patient may be incompletely served in descendent subproblems, but the incomplete service chains must be different.

Chain	Infeasibility
+2-->-1-->ASF_1	Missing node -2 (destination) between the patient's origin and ASF_1.
ASF2-->-3-->+3-->-2	Missing node +2 (origin) between ASF_2 and the patient's destination.

Figure 5.11. Incomplete service chains.

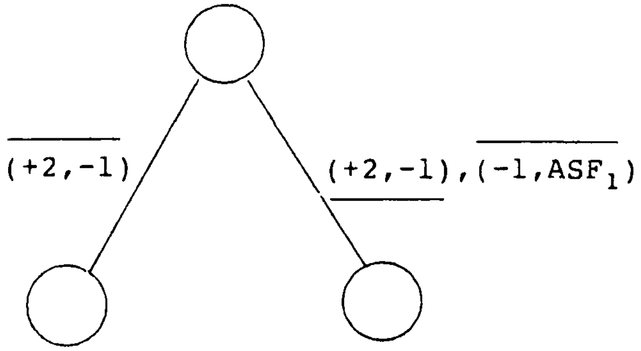


Figure 5.12. Incomplete service partitioning branching.

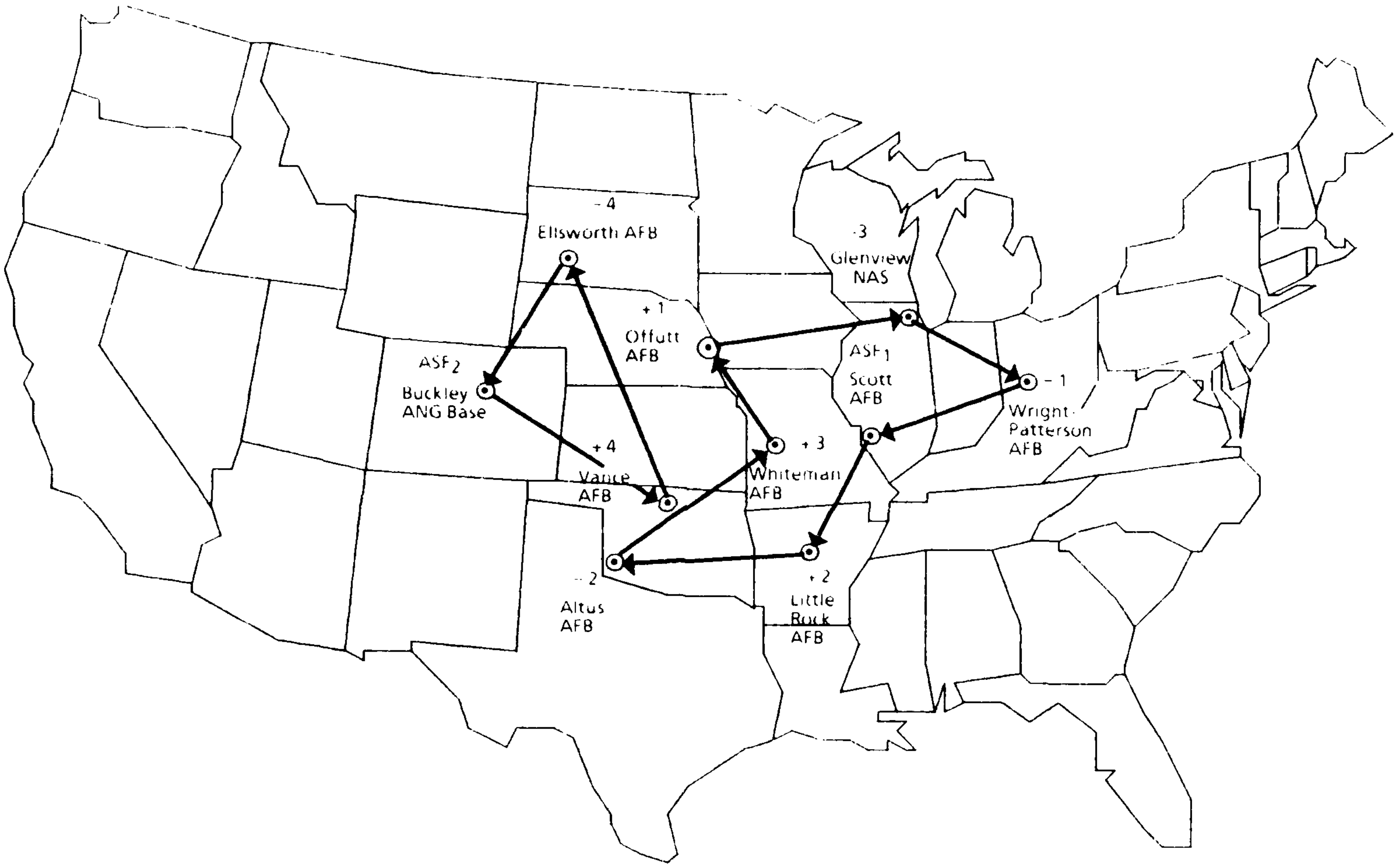
Precedence infeasibilities in Problem MDMRP, the second kind of infeasibility that can arise, are treated in exactly the same manner as in Problem SVMRP. The shortest infeasible chain in which an origin and destination are not in order is branched. For example, if the solution

$$ASF_1 \dashrightarrow +4 \dashrightarrow -3 \dashrightarrow +5 \dashrightarrow +3 \dashrightarrow -5 \dashrightarrow -4 \dashrightarrow ASF_1$$

$$ASF_2 \dashrightarrow +2 \dashrightarrow -2 \dashrightarrow +1 \dashrightarrow -1 \dashrightarrow ASF_2$$

occurs, in which patient 3's origin is visited after his destination, the infeasible chain ((-3,+5),(+5,+3)) would be branched. Both types of infeasibilities can appear in the same solution, in which case incomplete service branching should be used. (Order infeasibility cannot be demonstra‐ted in Figure 5.8 because a subtour cannot contain an infeasible chain if N is less than 3, by Lemma 4.)

5.5.6 An MDMRP Example. To demonstrate the MDMRP algorithm, consider the single aircraft problem used in the SVMRP computational demonstration in the last chapter (Figure 4.34). Let Scott AFB and Buckley ANGB be Staging Facilities 1 and 2 respectively, and designate Wright-Patterson AFB as the destination of patient 1. Figure 5.13 depicts the resulting MDMRP solution found by the MDMRP algorithm. Although the destination of patient 2 is closer to the route flown by the aircraft serving ASF_2, the cost of adding both the origin and destination of patient 2 to that route is substantially greater than the savings realized by deleting those stops from the route serving ASF_1. This reflects a major problem stemming from the assumption that routes only transit a single depot: the distribution of origin and destination points can cause aircraft routes to overlap regions. When the origins and destinations are in different regions, this procedure is ineffective, particularly when the regions are non-contiguous. As we will see in the next chapter, a different technique must be used to handle interregional transfers. Even then, this procedure can be used to find routes for adjacent intraregional problems, to check for situations where crossing boundaries can save travel distance.



OPTIMAL:2945 NM

Figure 5.13. Patient Locations, Problem IV.

5.4 Multi-depot, Multi-vehicle Routing Problems with Precedence Relationships (MDMVMRP). By combining the solution procedures for Problems MVRP and MDMRP, the Version III model can be solved. Specifically, Problem MDMVMRP requires (i) the creation of duplicate depots and a test to ensure that any subtour contains a copy of a depot to handle the multiple vehicle aspect, (ii) the restriction that any tour is infeasible, to satisfy the multiple depot routing feasibility stipulation, and (iii) the ordering restriction to satisfy precedence relationships. Figure 5.14 illustrates the idea that all three types of infeasibility that can occur in multiple depot, multiple vehicle problems with ordering precedence relationships. The situation labeled A is incomplete service to patient 2. Situation B is a violation of origin-destination ordering, and C contains a disconnected cycle.

In solving the MDMVMRP problem, ten different solution types can be found, not counting those special cases in which the dual objective of the MAP exceeds the upper bound before MAP optimality, or there are arc inclusion/exclusion conflicts. Table 5.6 shows the recommended branching[5] procedures for the various possibilities. Since a tour calls for one aircraft to visit all nodes, incomplete patient service in a tour is impossible by definition; therefore, a tour will fall into two cases only. And, since a tour is infeasible by definition, Table 5.6 calls

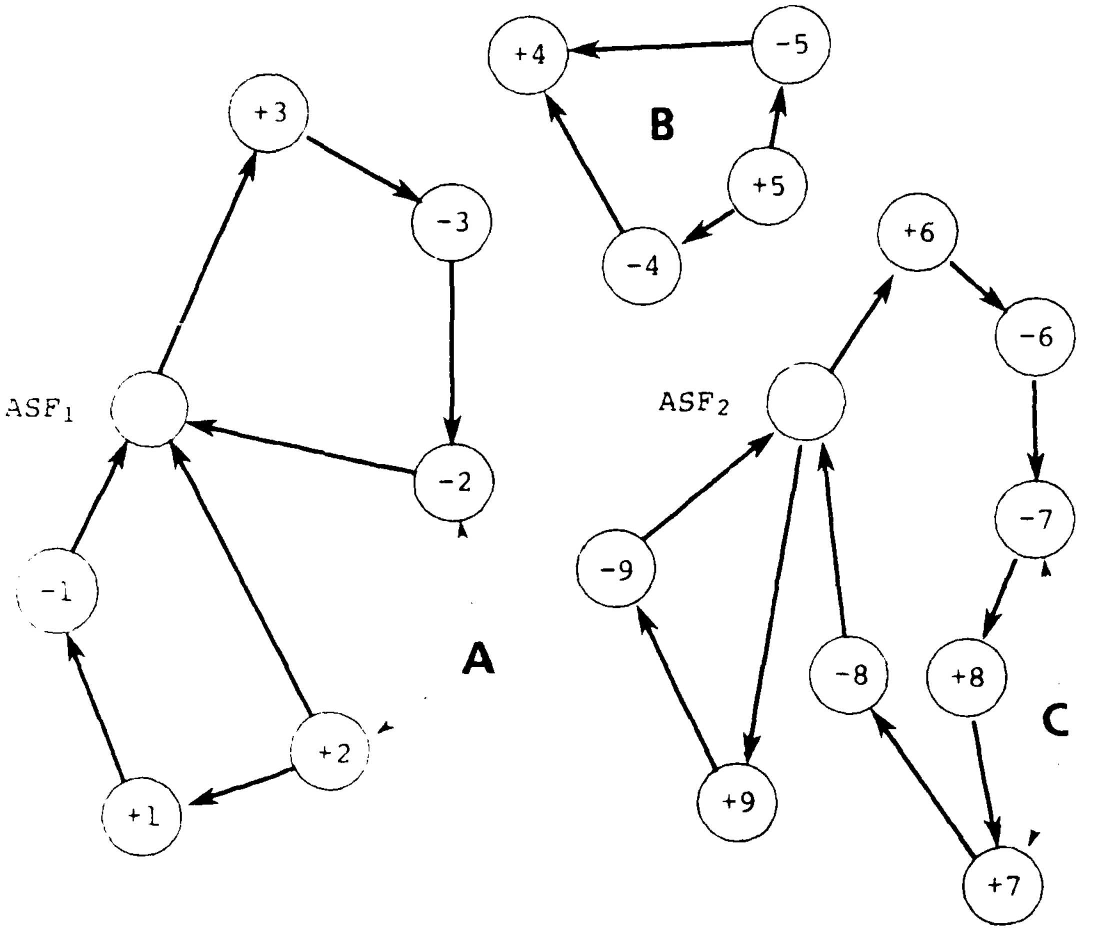


Figure 5.14. An MDMVMRP solution with three infeasibilities.

for the shortest infeasible chain to be branched, if one occurs in the tour, or to branch the entire tour otherwise.

Before we conclude our discussion of the MDMVMRP algorithm and Version III of the aeromedical planning model, we should address one other assumption we have made, that stops are either depots (staging facilities) or the unique origin or destination of a single patient. As we saw in the empirical data in the second chapter, airfields

TABLE 5.5
MDMVMRP Solution Procedures

Routing Solution Type	Precedence Infeasibility Resolution			
	Incomplete Service	Infeasible Chain	Both Infeasible Structures	Neither Infeasible Structure
Tour	Not Applicable	Branch Infeasible Chain	Not Applicable	Branch on Tour
Infeasible Subtour(s)	Branch Incomplete Service Chain	Branch Infeasible Chain, or Infeasible Subtour	Branch Incomplete Service Chain	Branch Shortest Infeasible Chain
Feasible Subtour(s)	Branch Incomplete Service Chain	Branch Shortest Infeasible Chain	Branch Incomplete Service Chain	Feasible

may be all three types of stops on the same day, particularly the staging facility bases.

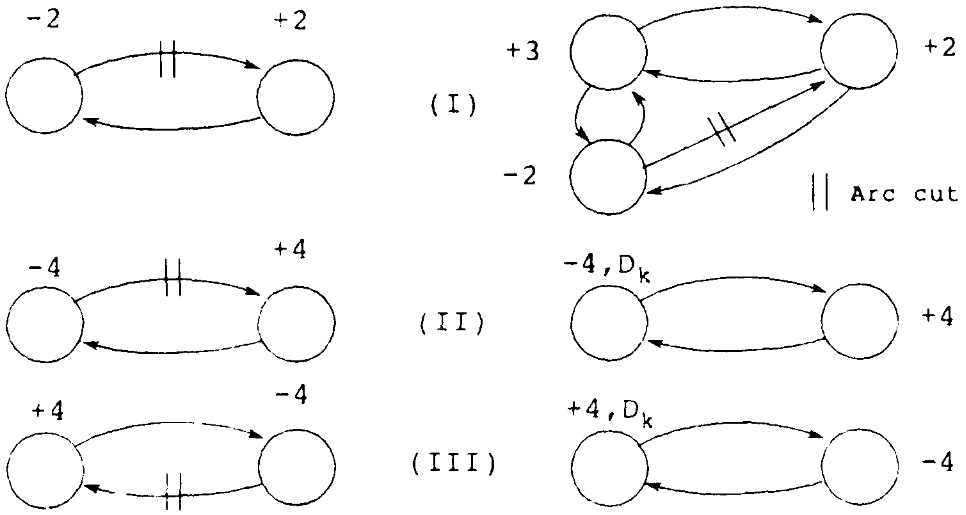


Figure 5.15. Non-unique points.

Three cases are shown in Figure 5.15. In the first, if one of two non-depot bases is both an origin and a destination, the node is replicated as shown, and arcs of zero length connect the two replicas. In the other two cases, when an origin or destination is also a depot, we cannot exclude arcs as we can when the depot is not an origin or destination. We could also use replication, but that is unnecessary if the infeasible chain test is modified to account for the fact that the depot is a service point.

5.5 Summary and Conclusions. We have shown that the single vehicle procedure can be extended to handle multiple vehicles, multiple depots and origin-destination precedence relationships. Tests have only been consistently successful on small problems (up to 50 nodes); in one using 84 bases, five depots, and a single aircraft at each depot, the algorithm would not converge in 300 seconds of CRAY-1S CPU time, although the optimality gap was less than 7 per cent. Additional research is needed to increase the limits of the method. Using the Tillman-Cain procedure and the modifications made by Golden, Magnanti, and Nguyen [GOLD77] to quickly generate good upper bounds is being explored.

As we noted, patient origins and destinations in different regions will limit the effectiveness of the MDMVMRP algorithm. However, it can still be useful in pointing out non-obvious instances where an aircraft from one region can serve a patient in another, when extending its route requires less travel time than using the aircraft serving that patient's region.

ENDNOTES

1. In virtually all asymmetric cost solution methods, symmetric costs can also be used. The reverse is generally not true. A good example is the method proposed by Gavish and Srikanth; symmetric costs allow them to use a derivative of an efficient greedy minimal spanning tree algorithm. Otherwise, they would have to use a less efficient minimal rooted spanning arborescence algorithm.

2. Other formulations treat m as variable. Bellmore and Hong [BELL74] use a fixed charge method to incorporate variable m, while others treat m as a parameter to be varied over different executions of a fixed m model.

3. Svestka and Huckfeldt claim that only 1/nth of the work is required to solve an n-city subproblem by using the final reduced costs of the parent problem optimal and new arc exclusions and inclusions, rather than the original costs. [SVES73] The new optimal is often found after one or two row and column reductions and an allocation step.

4. We relax this assumption in the next chapter.

5. If multiple infeasibilities occur, treating the one that results in the fewest branches is the rule of thumb we use.

CHAPTER VI

A MECHANISM FOR AEROMEDICAL TRANSPORTATION PLANNING

> *The night sky is a model of the cosmos that never, ever existed; all those points of light were generated at widely different times. But we use this model without even noticing, just as we hold up notions of religious or plotiical "truths" that are the products of models having a hopelessly low power to discriminate and hence to explain.*
> *Stafford Beer*

6.1 Introduction. In this chapter, we will extend the multiple depot, multiple aircraft, mixed service model by incorporating three additional characteristics of the DOD aeromedical transportation problem: incomplete or partial service (not visiting both a patient's origin and destination on the same route); regional organization; and routing restrictions. With these extensions, the Version IV model corresponds to specifications given in Chapter III. These properties are particularly difficult to include in the model because they greatly complicate the problem structure we were able to exploit in the first three versions. We will develop the model to help planners produce daily mission schedules that improve patient service, by reducing the time patients remain in the aeromedical system and routing costs. Using historical data, future weekly regional service can also be planned. The method has the troublesome aspect that it utilizes perfect knowledge of all demands for a week, and past data from non-optimal system behavior to plan optimal future behavior, but the purpose is to show insights into better opportunities for resource use from which planners can learn.

6.2 Modeling Three Additional Patient Movement Problems. In the last chapter, we assumed that aircraft were assigned to specific depots, that every aircraft mission began and ended at the same depot, and that missions did not transit other depots. We also assumed no limit to the number of flight segments in a single mission. And, we assumed that a transfer had to be completed on the same mission.

However, the actual patient flows we analyzed in Chapter II make these assumptions indefensible. Because of operating rules that impose a maximum crew duty period to avoid crew fatigue problems, the number of flight segments in a mission is limited. The number of missions each day is constrained by aircraft availability. On certain days, particularly Thursdays, more stops are required to meet every transfer demand than can possibly be made, even if the maximum allowable number of segments are scheduled.

Number of Distinct Service Points	Frequency
1-10	2
11-20	16
21-30	13
31-40	5
41-50	10
51-60	33
61-70	12
	91

Assuming that six aircraft are typically available and the maximum eight stops are scheduled on each mission, the number of service points that would have to be visited at least once to provide same-day service exceeded the maximum number of stops possible in over half the days we examined. (The maximum number of patient service points that can be visited is usually less than 48, when the movement requirements force different missions to visit some of the same points.) In this section, we will address the resulting problems of incomplete service, transfers, and the requirement to periodically rotate the aircraft through the central base for crew changes and aircraft maintenance.

6.2.1 Incomplete or Partial Service. Both the single and multiple depot models moved patients from their origins to the destination medical facilities on the same aircraft. Doing so implicitly assumed no restrictions on aircraft availability, route length, or maximum number of stops on any route. In effect, these were single period problems. With the operating restrictions given in our discussion of the system's environment, moving every patient on a single flight from origin to destination, providing what we have defined as complete service, is usually impossible.

Recognizing this, DOD stipulates maximum time criteria:

Patient Category	Required Pickup Response	Required Delivery Response
Routine	≤ 72 hours after movement requirement validation	≤ 72 hours after pickup
Priority	≤ 24 hours after validation	≤ 24 hours after pickup
Urgent	Immediate	Immediate

DOD intended that these response time "windows", between receipt of a movement requirement and patient pickup, and between pickup and delivery, allow flexibility and limit how long service could be deferred or delayed.

In effect, however, service time stipulations greatly complicate patient movement decisions. Figure 6.1 shows all possible decision combinations for routine patient movements that do not violate DOD guidelines. Five basic decisions are involved:

Label	Explanation
1. Defer Both Pickup and Delivery.	Defer pickup. (Patient remains at his origin until the following day.)
2. Pickup and Deliver.	Provide complete service.
3. Pickup, Defer Delivery.	Pickup patient and take to an ASF for first overnight stop.
4. Defer Delivery.	Leave patient at the same ASF or move to another ASF for another overnight stay.
5. Deliver.	Move patient from the ASF to his destination.

As often happens, the price of flexibility is increased complexity and reduced service quality. The twelve distinct paths in the decision tree are for one patient; decision makers typically handle from 20 to over 100 patients each day. The presence of alternatives may influence them to "satisfice", when a better schedule might afford more direct service for the patient. Eleven intermediate states in the decision tree require a delay at the patient's origin or overnight care at an ASF. The longest path could mean six days in the system.

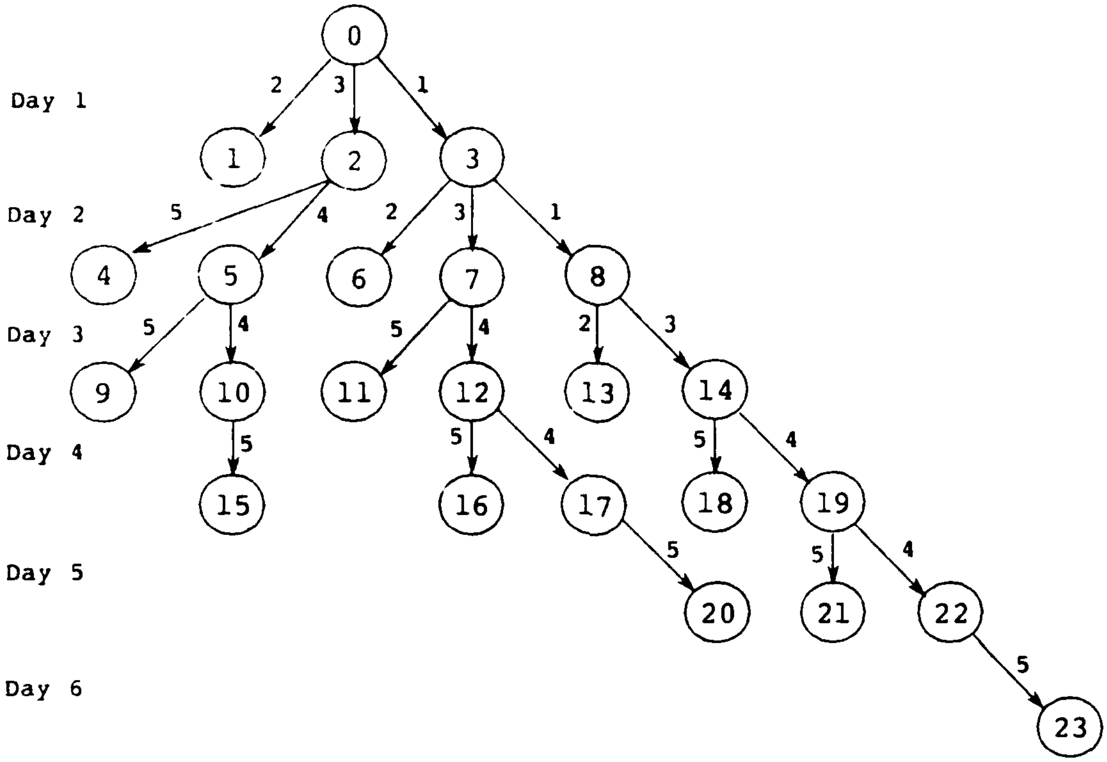


Figure 6.1. Patient service decisions.

6.2.2 Modeling Partial Service. The most important change introduced by decisions to defer pickup or delivery is the creation of multiple time periods. Let x_{ijk} = 1 if arc (i,j) is traveled in period k, and 0 otherwise, where i and j index patient origins, destinations, and ASF's. In the following illustration, let 1 index one patient's origin, 2 his destination, and day 1 be period 1. To extend our assignment problem-based formulation, let y_{ikl} = 1 if the patient's pickup is deferred from period k to period l, and

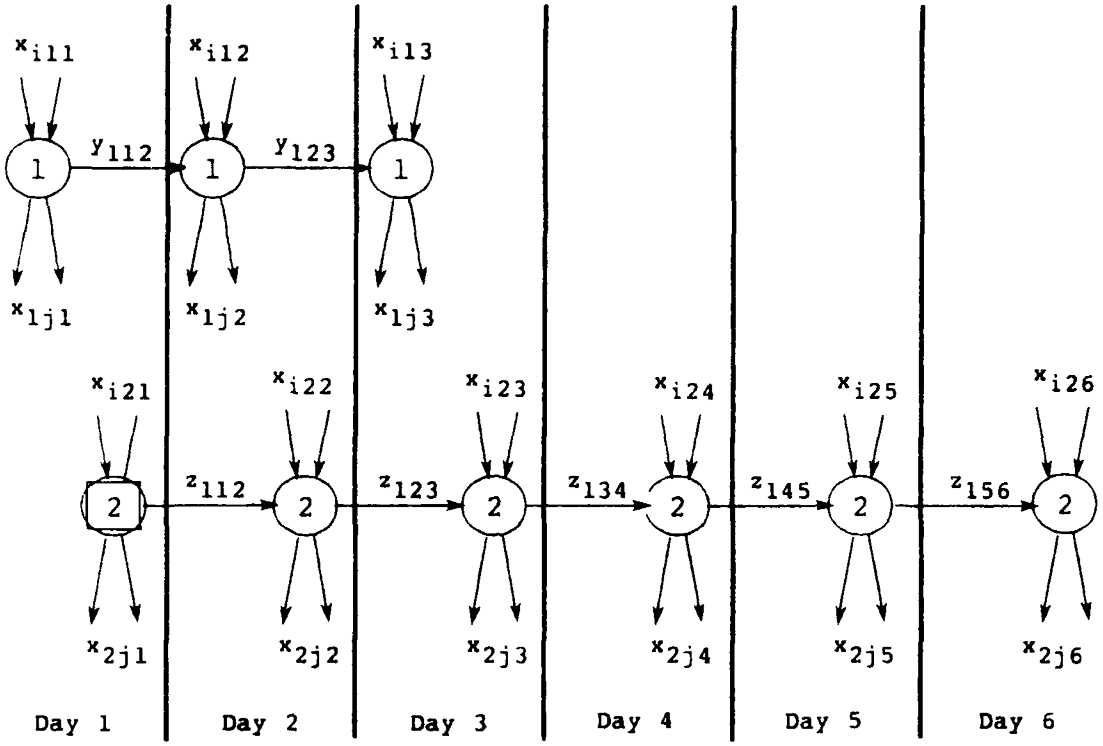


Figure 6.2. Routine patient service over time.

0 otherwise. Let $z_{ikl} = 1$ if the patient's delivery is deferred from period k to period l, and 0 otherwise. Figure 6.2 depicts the three types of decisions, where the vertically oriented arrows represent aircraft flight segments into and out of locations 1 and 2, and the horizontal arrows the two types of deferment decisions. That is, the flight segment variables represent routing decisions within the period, and the deferment decisions are the linkages across time periods.

We can now explicitly formulate the single patient, multi-period service problem. Find values of the variables x_{ijk}, y_{ikl}, and z_{ikl} that

$$\textit{Minimize} \sum_{i} \sum_{j} \sum_{k} \sum_{l} (\alpha y_{1kl} + \beta z_{1kl} + c_{ijk} x_{ijk}) \tag{6.0}$$

$$\textit{subject to:} \quad \sum_{i \neq 1} x_{i11} + y_{112} = 1 \tag{6.1}$$

$$\sum_{j \neq 1} x_{1j1} + y_{112} = 1 \tag{6.2}$$

$$\sum_{i \neq 1} x_{i21} + z_{112} = 1 \tag{6.3}$$

$$\sum_{j \neq 1} x_{2j1} + z_{112} = 1 \tag{6.4}$$

$$z_{112} \leq y_{112} \tag{6.5}$$

$$z_{123} \leq y_{123} \tag{6.6}$$

$$y_{123} + z_{123} \leq y_{112} + z_{112} \tag{6.7}$$

$$\sum_{i \neq 1} x_{i12} + y_{123} = y_{112} \tag{6.8}$$

$$\sum_{i \neq 1} x_{i1k} = \sum_{j \neq 1} x_{1jk} \qquad k = 2,3 \tag{6.9}$$

$$\sum_{i \neq 1} x_{i2k} = \sum_{j \neq 1} x_{2jk} \qquad k = 2,3,4,5,6 \tag{6.10}$$

$$\sum_{j \neq 1} x_{2j2} + z_{123} = z_{112} \tag{6.11}$$

$$\sum_{j \neq 1} x_{2j3} + z_{134} = z_{123} \tag{6.12}$$

$$\sum_{j \neq 1} x_{2j4} + z_{145} = z_{134} \tag{6.13}$$

$$\sum_{j \neq 1} x_{2j5} + z_{156} = z_{145} \tag{6.14}$$

$$z_{134} \leq z_{123} \tag{6.15}$$

$$\sum_{j \neq 1} x_{1j3} = y_{123} \tag{6.16}$$

$$\sum_{j \neq 1} x_{2j6} = z_{156} \tag{6.17}$$

where α and β are weights assigned to pickup delays and overnight stops respectively. Constraints (6.1) through (6.5) require that on day 1: patient 1 either be picked up or deferred (but not both) (constraints (6.1) and (6.2));

patient 1 should either be delivered or his delivery should be deferred (constraints (6.3) and (6.4)); and if his pickup is deferred, his delivery may be deferred, but not the converse (constraint (6.5)). Constraints for day 2, (6.6) through (6.11), for k=2, are similar, but also depend upon decisions made for day 1. For example, (6.8) requires that patient 1's pickup either be deferred or made in period 2 iff his pickup was deferred in period 1; if it was not deferred, the right hand side would be zero and both options in period 2 would automatically be zero. On day 3, the option to defer pickup is absent, so that constraints for days four through six refer only to delivery options.

Expanding the formulation to more than one patient would require the use of another summation in the objective, and a separate set of constraints for each patient. In addition, order and subtour elimination constraints would be needed to ensure route feasibility. With 22 constraints per patient, and 400-600 distinct patient movements per week, the number of constraints would be very large (12,320, assuming an average of 560 from the 90 day period we observed), even without the additional subtour and order restrictions. With this size, an integer LP (ILP) formulation of this problem would be impossible for current state-of-the-art computer codes to solve.

The assignment formulation is still incomplete, however, because it lacks a very critical linkage, between routes that pick up a patient in one period and deliver him in another. In the case of single period routes that provide complete service, the identity of the patient is not important. However, in multi-period problems, we have no explicit link between the end point of the route that picks a patient up and one that delivers him. We could test for the existence of a path between the end point of the former and the beginning of the latter. But, even if that connection exists, it may require the patient to remain in the system for an unecessarily prolonged time. The reason is that the assignment formulation will avoid deferments, with suitable choices of weights on the deferment variables, but since time spent in staging facilities is not explicitly modeled, that *time is not minimized.*

The problem is that the assignment formulation chooses the routing network, and not how the patients move (or flow) over network arcs. In single period problems, choosing only those arcs that force flows between arcs in correct order allows us to ignore indivdual patient flow, so we need not be concerned with patient identity. In multi-period problems, with patients needing more than one flight to reach their final destinations, we have to be concerned with patient identity, at least by groups with common origins or destinations, and use both routing design and flow constructs. After introducing two other complications, we will return to this issue.

6.2.3 Regional Organization and Aircraft Routing Restrictions. We introduce these complications together because they are closely related. With multiple regions and a significant number of interregional transfers, the restriction that aircraft begin and end missions at the same staging facility changes to the stipulation that the end points be any staging facility, but not necessarily the same one. We then have another problem to resolve, that with one central base for aircraft maintenance and crew basing, we need to ensure that the aircraft periodically cycle through that base. (The unwritten rule is once every one to two days.) Finally, each mission is limited to eight stops, unless bad weather reduces that figure to seven. And, the number of missions is restricted to seven, with six desirable.

Figure 6.3 illustrates a simplified multi-regional problem with seven type of patient transfers, based on (1) whether the origin and destination are in the same regions and (2) node type (origin or destination). The attributes of each type are given in Table 6.1. One transfer (#7) involves non-adjacent regions, which could create a separate category. However, this rarely occurred in the actual data; most transfers were like those of patients #5 and #6.

Assuming that the maximum number of stops is six, then if depot #1 is the starting point, the aircraft stopped at

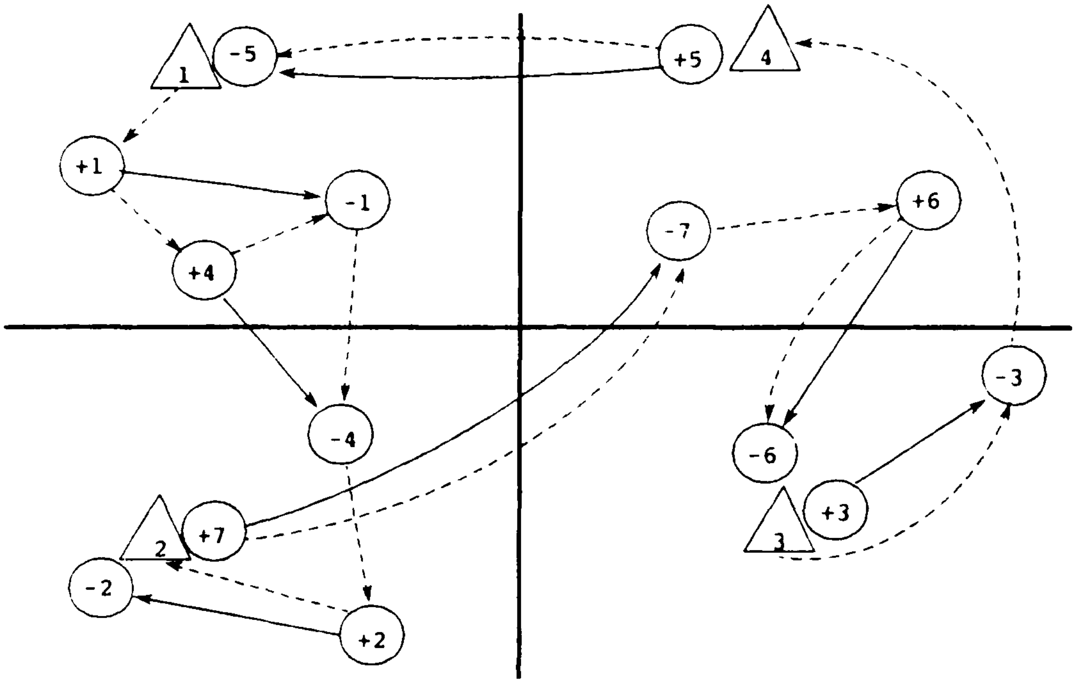


Figure 6.3. Multi-regional patient service.

Depot 2 after six flight segments. On the second day, six stops were again made, and the aircraft and crew returned to the central base. This example meets the desired goals of using all but not more than the maximum number of stops on each mission, and returning to the central base periodically (two to three days typically).

We have not seen this problem in the routing and distribution literature. Lokin [LOKI78] developed procedures to solve traveling salesman problems with stops occuring in clusters that must be served contiguously before other other clusters, and Cullen, Jarvis and Ratliff [CULL81] use center-of-mass techniques to find origin and destination clusters. Only Reufli [RUEF71] explicitly

TABLE 6.1

MULTI-REGIONAL PATIENT TRANSFERS

Patient Number	Transfer Node Locations	Node Types	
		Origin	Destination
1	Same Region	Non-Depot	Non-Depot
2		Non-Depot	Depot
3		Depot	Non-Depot
4	Different Regions	Non-Depot	Non-Depot
5		Depot	Depot
6		Non-Depot	Depot
7		Depot	Non-Depot

considers regional organization, and then only in a fixed transshipment network context in which routing is not considered. Regional organization could be a natural way to decompose problems, but in the aeromedical system, more than half of all transfers are made across regional boundaries, which greatly reduces subproblem separability.

To accomodate the aircraft routing limits of maximum route length, an assignment-based formulation can be changed by adding a restriction that the sum of all arcs on a route be less than the maximum. However, we have not explicitly represented the concept of a misssion. Bodin et al [BODI81] do this by using variables x_{ij}^{k}, where k is the index of aircraft k. Route length limits convert problems into the general vehicle routing problem. Christofides et al [CHRI81] claim that the largest problems solved exactly with such additional constraints have only 25-30 stops and a single depot. Therefore, it seems unlikely that assignment methods will fare well with up to 7 periods, 6 depots, and up to 100 stops in each period.

6.3 Methods for Solving Multiperiod Aeromedical Routing Problems. At least two research efforts have addressed the multi-period problem. We discussed Swoveland's multi-period, multicommodity, production-distribution research in the last chapter. [SWOV71] His problem did not include routing concerns. Russell and Igo [RUSS77] devised heuristic methods to assign service points to specific days of the week, and then used single period routing methods to find routes. In their formulation, the number of times a point is served is fixed in advance, the time between visits can be specified, and specific or permissible service days can be stipulated. They devised one heuristic to cluster points with the same frequency based on proximity, and two heuristics to create routes through clusters. One uses multi-depot exchange procedures based on Lin's k-opt method [LIN 73] for problems with up to 300 stops, and the second a modified Clarke-Wright savings procedure for larger problems. Russell and Igo were only concerned with pure delivery problems, and not the many-to-many case.

The most promising approach is to consider network routing design and individual patient flow planning simultaneously. Currently, problems of the size involved in finding a weekly optimal schedule that minimizes the combination of patient travel cost (total individual travel time) and network operating cost (total aircraft time or distance) are beyond the capacity of present algorithms. However, parallel efforts to develop efficient routing and multicommodity flow models, using the resource-directive decomposition technique described earlier, coupled with supercomputer capacity and speed, may realize that goal.

The fundamental change to our assignment-based formulation is to consider patients with a common destination as a commodity. (We use destinations because of the concentration of flows from relatively more origins to fewer destinations we observed in Chapter II). As before, let N be the service points in all demands for one week; the same point on a different day is notionally a different point. Also, assume we have replicated the depots (staging facilities) and nodes that are both origins and destinations for different patients. Rather than allow all arcs between all nodes, let the set of arcs A consist of all connections between nodes in the same period, provided arcs are less than the maximum aircraft range (assume 1800 nautical miles), and do not directly connect a patient's destination with his origin. Let j represent the arc (i_1,i_2), in standard node-arc incidence notation.

Let K be the number of different commodities (patient destinations). For a weekly problem, the 54 bases that explained over 99 per cent of all origin-destination pairings would adequately represent all service points. To evaluate additional stops (e.g., the 21 not currently served that planners have been requested to add), that number could be increased, but the number of commodities directly affects problem size and must be limited to the minimum necessary to represent the movement problem. Let r_i^k (> 0) be the number of patients to be moved from node i to the destination k; let $\underline{r}^k$ be the vector of movement demands from all origins to destination k. ($|\underline{r}^k|$ would be the demand at node k.) If we do not visit both the origin and destination nodes on the same day, thereby deferring service, we need to create deferment arcs from those nodes to replicates in subsequent time periods (creating them if they are not in the movement requirement node set for those days; as transshipment points, they would have neither a supply of nor demand for commodity k). Movement precedence represented by not allowing priority patient origins or destinations deferment arcs. Because urgent patients are not routinely scheduled by <u>advance request</u>, we are not considering them in daily and weekly route planning.

Let c_j be the aircraft flight time between nodes i and j, x_j the decision to include arc j in the routing network, f_j^k the flow of commodity k over arc j, and F_j the flow cost of one unit over arc j. For the deferment arcs, F_j could be set to, say, 12.0, representing an overnight stay at an ASF or a one-day pickup deferment. One important issue we have not yet raised is aircraft capacity; let U_j be the total capacity of arc j for all commodities. In our problem, U_j typically is the maximum capacity of the C-9 aircraft, which is 40 patients. There are other capacities that we will describe later.

The general multicommodity problem formulation is given as follows:

$$\text{Minimize} \sum_{j} \sum_{k} F_j^k f_j^k + \sum_{j} c_j x_j \qquad (6.18)$$

$$\text{subject to} \qquad A f^k = r^k \qquad (6.19)$$

$$\sum_{k=1}^{N} f_j^k = U_j , \quad j \in A \qquad (6.20)$$

$$f^k \geq 0, \qquad (6.21)$$

where constraint (6.19) requires that all commodity movement requirements be met, (6.20) restricts arc flow to a specified limit, and (6.21) ensures that all flows are strictly non-negative. Magnanti and Wong [MAGN84] describe this linear multicommodity minimum cost flow problem with fixed charges as one combining both network design (in the choice of variable x_j values) and optimal commodity flow. As such, it has the potential to incorporate the concerns of both patients and those who pay for their transportation costs, since it will minimize both aggregate travel time and total routing cost.

Success in solving multicommodity problems is limited at best, and even less so in the fixed charge case. But, of particular interest to us here is the work of Ali et al [ALI 81] that we described in Chapter III. They attack the problem by solving it in two phases. First, they eliminate the fixed charge portion of the problem, and solve a simpler multicommodity minimum cost network flow problem to which the following additional constraints have been added:

$$\sum_{k=1}^{N} x_j^k \leq y \qquad (6.22)$$

$$Ay = 0, \qquad (6.23)$$

which requires that flows occur in circuits. In our problem a circuit can be interpreted as a route. For that reason, they call this a route generator model. Now, circuit length and aircraft capacity are excluded from this formulation, so the resulting solution, which they decompose into new routes that have no arcs exceeding aircraft capacity, is not feasible without the decomposition. However, decomposition does not eliminate excessive route length in the process of finding these "nominal" routes.

To do this, they ask planners to consider the nominal routes generated by their route generator model, and revise them as necessary to create feasible and acceptable routing alternatives. Undoubtedly, this affords planners the opportunity to introduce other concerns that cannot be modeled easily, and use their discrimination and ability to eliminate alternatives that the mathematical model cannot.

With these revised routes as candidates, they then solve the fixed charge problem, with an important change. Instead of allowing all arcs in the networks to be candidates for inclusion in the network, they restrict the arc set to those in the candidate route set. In fact, their route selector model replaces the fixed charge portion of the objective (6.18) with the expression

$$\sum_{l=1}^{L} c_l v_l ,$$

where v_1 is a binary variable representing the inclusion or exclusion of route R_1. By adding the expression

$$\sum_{j \in R_l} \sum_{k=1}^{N} x_l^k \leq M^* v_l , \quad l=1,\ldots,L \tag{6.24}$$

they force the variable v_1 to assume a value of 1 if the route is used. In other words, they choose entire routes, and not individual arcs, in this second phase, finding the optimal set of routes that produces the lowest possible combination of routing and patient flow costs.

There are a number of features that are important to the aeromedical problem. First, the explicit incorporation of both measures of performance is in keeping with our original goal to satisfy the needs of two client groups. Secondly, it resolves the problem of patient identity that the assignment formulation could not. Third, in their approach, planners intervene to revise the nominal routes before final route selection, which removes the "veto-only" arrangement for which Mason criticizes many LP decision models. And, they include several other constraints, such as cycling through a home base, that are in our specifications in Chapter III.

To see how we might modify their model to solve the aeromedical weekly routing problem, consider the following depiction. Here we have included the depots, denoted by triangles, as the beginning and ending points of routes through the demand points for a day. The arc between them, across the line separating the days, represents the staging facilities, and the upper bound U_j on that arc would be the

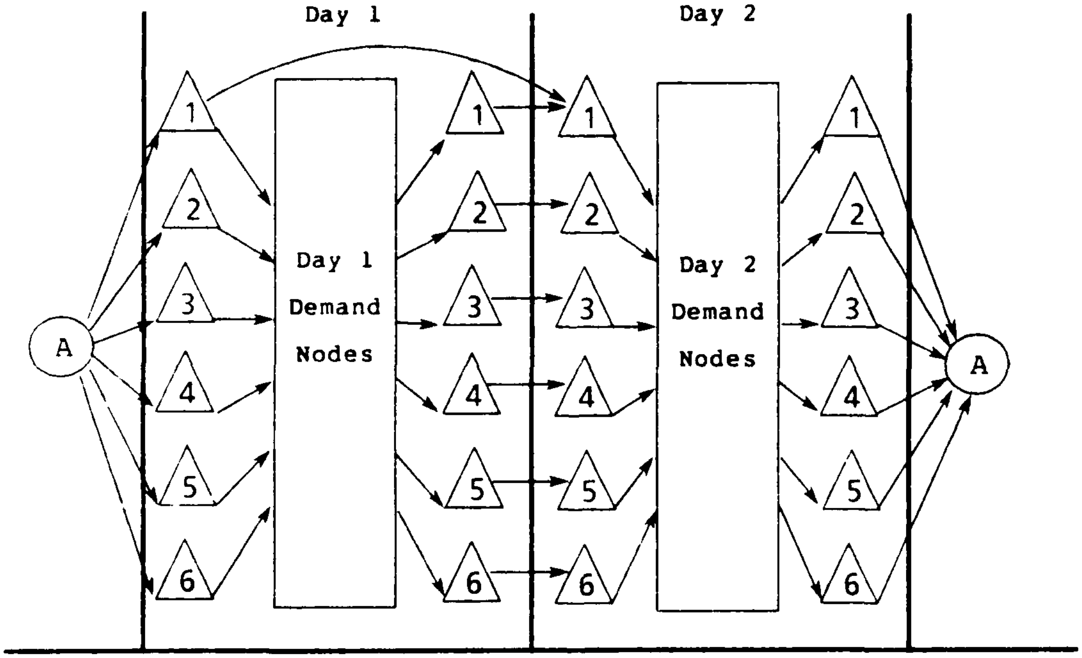


Figure 6.4. Two-period patient service.

ASF capacity. The nodes labeled A are a special commodity: the aircraft. Using this artifice, we can vary the supply of aircraft for the week, and also each day, by including an aircraft arc between each ASF node. By specifying that all aircraft arcs between ASF's over the end of a day belong to a bundle constraint, we can also impose daily availability limits. The arc from the first copy of ASF 1 on day 1 to the corresponding node in the second day represents an aircraft not used.

Our contribution to the eventual solution of this model is a specialized route generator. As Figure 6.4 implies, multi-period problems are nearly separable into a series of single day subproblems, linked by deferments, flows through the staging facilities at night, and aircraft availability. The object of routing is to both minimize aircraft travel (which also reduces patient travel time in some instances) and aggregate patient travel time. With deferment weights large relative to the elapsed time of same-day delivery, the best route set should provide complete service to the largest possible number of patients, over the shortest feasible route set.

Ali et al [ALI 81] use a route generator that attempts to greatly reduce the number of candidate arcs by allowing only five arcs to be incident out of a node. Depot nodes completely interconnected. The routes generated by the multicommodity flow and circuit decomposition routines produce routes that are feasible in flows, but possibly infeasible in route length. We have developed an alternative method that produces routes that are minimum in arc length, no longer than the maximum number of segments, and easily adapted to planner interaction and participation.

The essence of our procedure is merge a heuristic set partitioning technique developed by Cullen, Jarvis and Ratliff [CULL81] with our SVMRP algorithm and other techniques for generating the routes in the object set. Any routes, including those generated by Ali et al vector-circuit decomposition, can be columns in the set partitioning model, which is given as follows. Find a set of routes $J^k = \{j | v_j{}^k=1\}$, called a partition, that will

$$\text{Minimize} \sum_{j=1}^{n} c_j v_j \qquad (6.25)$$

$$\text{subject to:} \sum_{j=1}^{n} a_{ij} v_j = 1 , \quad i = 1,2,\ldots,m \qquad (6.26)$$

$$v_j \in \{0,1\}, \qquad (6.27)$$

where m is the number of movement origin-destination (O-D) pairs, n the number of routes, and $a_{ij} = 1$ if both points in O-D pair i are in route j. The solution $V^k = \{v_1, v_2, \ldots v_n\}$ is the vector of binary variables in which v_j is one if it is in the partition, and 0 otherwise. Cullen et al define row prices $P^k = (p_1{}^k, p_2{}^k, \ldots, p_m{}^k)$ as a feasible row price (dual variable) vector for the m demand constraints of the partition J^k if

$$\sum_{j=1}^{n} p_i^k a_{ij} = c_j \ , \ j \in J^k . \tag{6.28}$$

They interpret p_i^k as the estimated cost of serving O-D pair i. They then prove that for some partition J^1 and price vector P^1, any other partition J^2 has value

$$Z^2 = Z^1 - \sum_{j \in J^k} \sum_{i=1}^{m} (p_i^k a_{ij} - c_j) . \tag{6.29}$$

The solution V1 is optimum if for any P^1, and partition J^1,

$$\sum_{i=1}^{m} (p_i^k a_{ij} - c_j) \leq 0, \ j \in J^1, \tag{6.30}$$

the quantity they call the <u>potential</u> <u>savings</u> over Z^1 that would result from using another partition that includes column j. Of course, only columns with nonnegative savings would be included in the new partition.

There are three basic procedures in the partitioning technique. The first is row pricing. A feasible starting route set can be generated by finding the shortest route to serve each O-D pair separately. To do this assign the closest depot to the origin first, then the origin. If the distance between origin and destination is beyond the maximum range of the aircraft, then find the shortest path from the origin to the destination via depots only. Otherwise, use the great circle distance. Then, assign the destination and the closest depot to it. The rationale for the shortest path via depots is that route subsets that end at depots provide incomplete, but feasible, service.

To find a price vector for routes in a new partition that contain several multiple O-D pairs, use the following proportional pricing scheme. Suppose route R_8 with length c_8 serves patients 1, 3, and 4. Single patient routes cost c_1, c_3 and c_4. Observing that

$$c_8 = c_8 \frac{(c_1 + c_3 + c_4)}{(c_1 + c_3 + c_4)} = \frac{c_1 c_8}{(c_1 + c_3 + c_4)} + \frac{c_3 c_8}{(c_1 + c_3 + c_4)} + \frac{c_4 c_8}{(c_1 + c_3 + c_4)}, \qquad (6.31)$$

let the three terms on the right hand side be p_1^k, p_3^k, and p_4^k respectively.

Pricing is necessary whenever a new partition is found, which is accomplished as follows:

Step 0: Let $J^2 = \varnothing$ (J^2 will be the indices of columns in the new partition) and $N=\{1,2,...,n\}$, (N will be the indices of columns which are candidates for inclusion in J^2)

Step 1: Calculate the potential savings (6.30) for $j=1,2,...,n$.

Step 2: Pick the column k in N with the largest potential savings.

Step 3: For $i=1,2,...,n$ if $a_{ik} = 1$ set $a_{ik} = 0$ for all $j \neq k$. Note ... that since any subroute of a feasible route is also a feasible route, the new columns are legitimate.

Step 4: Let $J^2 = J^2 \cup \{k\}$ (i.e., put column k in the new partition) and $N = N - \{k\}$.

Step 5: Delete from N all j for which $a_{ij} = 0$ for all $i+1,2,...,m$.

Step 5: If $N = \varnothing$ stop. Otherwise go to step 2.

[CULL81,p.128]

This procedure can create new routes by eliminating one or more demand pairs from an existing column (route).

Rather than use a heuristic such as Clarke-Wright to find the routes and route costs c_j, we employ the SVMRP algorithm for both intraregional and interregional sets of O-D pairs. With a current price vector and costs c_j, the heuristic partitioning procedure finds a new partition, if the current one is not optimal. The pricing procedure is then run, and the process continues until optimality is reached, a specified number of partitions is generated, or some other stopping condition is met.

Route generation involves several procedures. The first finds the single patient routes with which to create the first feaible row price vector. SVMRP is then called to combine all points with origins and destinations in the same regions into intraregional tours and interregional paths. Next we find instances where the origins and destinations of interregional O-D pairs can be served by intraregional routes, and where routes can be concatenated (joined) without exceeding maximum allowable route length. Finally, in roughly 20 per cent of the demand pairs, we observed two routes, e.g., ADW ==> PVD ==> ADW and ADW ==> PVD ==> ORF ==> ADW, where one route subsumes the other.

Typical performance of this procedure is to generate a feasible route set of 6-7 routes that will provide 80 per cent complete service for 75 demand pairs, and over 90 per cent complete service for 40 pairs or less. On weekends, when the demand is usually under 30 pairs, and when most originations are overseas patients being moved to hospitals in the US from either Travis or Andrews Air Force Base, the route set satisfies all demands. As a final note, we speculate that good initial routes will provide the resource-directive multicommodity algorithm with an advanced starting basis that will improve the algorithm's performance.

This procedure has worked well enough to warrant further tests and extensions that bring planners into the route generation and selection process. Cullen et al suggest that the combination of computer-generated graphics displays and route planner interaction improves the quality of routes generated, and perhaps more importantly, increases planner involvement. , As we discussed earlier, the organization currently lacks the necessary computing equipment for this.

The use of the multicommodity fixed charge network flow approach will require further developments of fixed charge algorithms. Until that time, we can run the multicommodity problem with a fixed route st icture given by our routing procedure. A parallel research effort by Kennington, to solve the wartime patient transportation problem, is currently developing a resource-directive decomposition technique to solve the multicommodity flow problem. Recalling that each of the 54 commodities (patients with the same destination) compete for aircraft capacity, that resource is allocated by the master problem so that (1) solutions are feasible in arc capacity and (2) the subproblems are pure minimum cost network flow problems that can be solved efficiently.

Our collaborative efforts with Dr Kennington involve another potentially significant factor in eventually solving the route selection problem optimally. the use of supercomputer processing speed, vector processing capability, and large core capacity. Problem size has been severely limited and computation times excessive in previous efforts involving *even moderate size* problems. The multicommodity fixed charge algorithm took 23 hours of cpu time, without reaching optimality, on a 60-node problem roughly equivalent to a single day aeromedical planning problem. The CRAY-1S on which our routing procedures were run, and the Control Data Corporation CYBERNET CYBER 205, on which both Kennington's resource-directive procedure and our routing model will be run, are faster by several orders of magnitude than the CYBER 73 on which Kennington ran the LOGAIR problem. And, he reports in a private communication that his experimental resource-directive code is dramatically faster than both primal partitioning and price-directive codes, for small numbers of commodities.

cent complete service for 75 demand pairs, and over 90 per cent complete service for 40 pairs or less. On weekends, when the demand is usually under 30 pairs, and when most originations are overseas patients being moved to hospitals in the US from either Travis or Andrews Air Force Base, the route set satisfies all demands. As a final note, we speculate that good initial routes will provide the resource-directive multicommodity algorithm with an advanced starting basis that will improve the algorithm's performance.

This procedure has worked well enough to warrant further tests and extensions that bring planners into the route generation and selection process. Cullen et al suggest that the combination of computer-generated graphics displays and route planner interaction improves the quality of routes generated, and perhaps more importantly, increases planner involvement. , As we discussed earlier, the organization currently lacks the necessary computing equipment for this.

The use of the multicommodity fixed charge network flow approach will require further developments of fixed charge algorithms. Until that time, we can run the multicommodity problem with a fixed route structure given by our routing procedure. A parallel research effort by Kennington, to solve the wartime patient transportation problem, is currently developing a resource-directive decomposition technique to solve the multicommodity flow problem. Recalling that each of the 54 commodities (patients with the same destination) compete for aircraft capacity, that resource is allocated by the master problem so that (1) solutions are feasible in arc capacity and (2) the subproblems are pure minimum cost network flow problems that can be solved efficiently.

Our collaborative efforts with Dr Kennington involve another potentially significant factor in eventually solving the route selection problem optimally: the use of supercomputer processing speed, vector processing capability, and large core capacity. Problem size has been severely limited and computation times excessive in previous efforts involving even moderate size problems. The multicommodity fixed charge algorithm took 23 hours of cpu time, without reaching optimality, on a 60-node problem roughly equivalent to a single day aeromedical planning problem. The CRAY-1S on which our routing procedures were run, and the Control Data Corporation CYBERNET CYBER 205, on which both Kennington's resource-directive procedure and our routing model will be run, are faster by several orders of magnitude than the CYBER 73 on which Kennington ran the LOGAIR problem. And, he reports in a private communication that his experimental resource-directive code is dramatically faster than both primal partitioning and price-directive codes, for small numbers of commodities.

6.4 Results and Conclusions. In this chapter we have formulated the full aeromedical planning model to include additional complications caused by partial service, regional transfers, and routing and aircraft operating limits. The resulting model cannot be solved by the assignment techniques used in the first three versions because the identity of patients is lost when they must be served by more than one mission. A fixed charge linear multicommodity minimum cost network flow formulation incorporates both patient flow and aircraft routing problem characteristics, but existing solution methods cannot solve programs as large as weekly aeromedical planning problem.

The approach taken by Ali et al, to first generate feasible route sets, and use those routes to find the optimum network design and patient flow, appears the most promising. Our contribution to eventually solving the full problem optimally is to incorporate our single vehicle many-to-many routing model into a set partitioning model that generates good and even optimal route partitions, and good advanced starting bases. With further refinements of this technique, and the estimate that a large-scale resource-directive code will be available within one year, the prospects for complete solution of the weekly scheduling problem appear excellent.

CHAPTER VII

CONCLUSIONS AND RECOMMENDATIONS FOR FUTURE RESEARCH

In the thesis we have examined the design of a planning system for the US Department of Defense aeromedical transportation system. We have found that that system is both complex in its own right, and yet is still only part of a much larger health care delivery system serving several million beneficiaries. We have been primarily interested in two aspects of the planning system, the planning decisions that must be made and the information required to make them. Our approach was to devise a framework that would incorporate both concerns, and then attempt to construct the actual planning mechanism within that framework. And attempt we did, as we discovered that our ideal of a developing a planning system that would produce optimal decisions for using the system's resources to the best advantage of its clients was ultimately unattainable. However, we also learned something about why that happened, and what is needed to overcome the problems we encountered. With the research currently underway, the prospects are good for resolving them. In the process, we did achieve a number of goals, and we observed a number of areas that warrant further attention.

We first examined the philosophical and epistemological nature of planning systems. Our basic conclusion was that there is no scientific foundation for the selection of any particular framework for planning system design. Any choice depends upon many strong systemic judgements. Perhaps most importantly, we concluded that our two most important concerns should be what we eventually wanted to implement, in terms of improving the condition of the clients the system serves, and how to guarantee that improvement actually results, and that disaster does not. The unavoidable conclusion is that the system's guarantor must always be its own participants, its planners, decision makers, and clients, whose intentions are perhaps the most crucial aspect of the whole system.

We chose as the principal clients the patients the system moves between medical facilities, and the public who provide the resources and share in the benefits of the system's performance. We first examined the larger Military Health Services System, and how it provided health care to its beneficiaries. As a number of studies have shown, its convoluted structure and lack of explicit health benefit program assignments created a number of problems that the transportation system has to contend with, including such things as a strong same-service bias in treatment referrals. We found its most significant shortcomings when we attempted to design an ideal structure for health care delivery within which the transportation subsystem would operate. Patients can be treated under a variety of alternative programs, and there is almost a universal lack of adequate data on care costs, beneficiary identification and other data critical to assessing care alternatives, including transportation. The system has been examined critically for ways to improve it, but unfortunately most previous efforts have utilized an efficiency approach in which cost reduction is the principal (or only) measure and benefits are either not calculated or simply ignored.

We also looked at the impact of the two major roles of the system and the way they both conflict with and complement each other. One of the principal missions of the aeromedical system is wartime casualty distribution from battlefield to hospitals, including those in the US. Budget support for the whole program is based upon wartime training, and yet the domestic system has never operated in its wartime role, beyond exercise and rehearsal, since World War II, and the principal measures of performance used by system managers are expressed in terms of peacetime performance. Many of the reasons for patient travel are directly attributable to differencesin the two roles.

After examining the system in terms of the nine categories devised by Churchman to describe and explain systems, we then used historical data to better understand how patients are transported by the system. We used statistical factor analysis techniques to extract significant structural properties of patient flows and the aircraft route structure. The dominance of intraregional sub-system patient flows and flows into and between a few major medical centers was clearly evident. Also evident was the relative concentration of demands on certain days of the week, such that on most days, more stops would be required to provide complete service to all patients than the system can make. This directly causes the intractable problems we described in the last chapter.

We then examined mathematical programming methods that incorporate movement characteristics such as multiple time periods and regionalized structure. One technique, the use of resource direction in decomposing large mathematical models, allows us to incorporate temporal and structural features, and provides a framework for integrating information flows and decisions. Resources are allocated to decomposed portions of the whole problem, say to a time period, a patient category, or regional unit, and the smaller problem is solved. These solutions, along with prices giving the value of the resources allocated, are returned to the central coordinating program, which uses that information to improve the resource allocation.

We then examined the basic routing problem and developed a series of increasingly more complex routing decision models. The first of these, the single vehicle, many-to-many routing algorithm, solves the problem of routing a single aircraft to serve patients with both origins and destinations. The SVMRP algorithm will solve problems larger than those developed for dial-a-ride applications, but does not handle time windows. Origins and destinations need not be unique, and the depot may be either an origin, a destination, or both. The assignment-based formulation we used seems quite efficient, particularly when compared with traveling salesman problems (to which it is related) of comparable size.

We then extended the single depot, single aircraft model to allow multiple aircraft, multiple depots, and origin-to-destination service requirements. The extension was relatively straightforward. The MDMVMRP model is of limited use in problems where the movement demands are not strongly or entirely intraregional. Also, maximum problem size is limited currently to 50 stops or less; increasing that limit should be the subject of future research.

Finally, we added the additional complications of partial service, interregional transfers, and aircraft operating restrictions, such as maximum stops per mission, maximum periods without visitng the central base, and starting and ending points at staging facility bases only. We immediately discovered that our assignment-based formulation could not be extended, but rather had to be used for generating routes for the fixed charge multicommodity minimum cost network flow formulation of the Version IV model. Current state-of-the art in resource-directive methods for solving the flow model will not handle entire weekly aeromedical planning problems, but current algorithm research indicates that with supercomputer processing speed and capacity, problem size limits can be expected to increase significantly very soon. Our route generator model provides good solutions to weekly problems, and an advanced starting basis for the flow model.

We have a number of recommendations for improving the planning process through the adoption of better planning technologies. In some cases, lack of data is a major problem. But in others, particularly in flight planning, lack of integrated data sources on weather, airfields, air route structures, navigation aids, and aircraft performance, not a lack of data, means that for all the data available, it cannot be directly accessed by computer-based planning models. A significant portion of Air Force and DOD planning data, for example, is only available in hard copy.

Despite extensive patient data collection, there is remarkably little use made of it for modeling patient movement. In addition to adopting the factor-analytic techniques we used, the wing should strongly consider using logit analysis with diagnostic data and other techniques to obtain insights and into and perhaps even predict movement workloads and trends. Better historical movement records that include more than just final transactions would be particularly useful. Another posible use of movement data is to determine if and how patient regulation could be improved. As we noted in an example, the destination choices made by movement regulators directly determine the of movement structure. This is an important area for future research.

We identified a number of technical issues in the development of the various versions of the route planning model that future research should address. We identified two measures of performance, network design costs (aircraft operating cost over network arcs) and flow costs (patient travel time). We were able to include them both only by creating a composite objective. Recent advances in multiple objective mathematical programming may permit the simultaneous consideration of separate objectives.

One needed extension to the routing model is to be able to handle dynamic demands. Our model builds routes based on advance requests. A significant number of urgent category patients require immediate movement, which requires the wing to either launch an unscheduled airplane or reroute one already assigned to a route. Dial-A-Ride solution methods of handling time windows, a problem that aeromedical planners sometimes encounter, should also be examined for possible inclusion in the routing model.

In our discussion of system objectives, we observed that one of the benefits of adopting effective planning models is that in wartime, when we would expect movement requirements to be extremely high and time frames greatly compressed, sophisticated methods would be need to effectively utilize the system's resources. A very important question is whether routing methods that are effective in peacetime are effective or even appropriate for wartime planning. It is not clear, from current research in this area, what the wartime movement problem might even be like.

Finally, in the true spirit of ideal planning, there are a number of "givens" that should be challenged. Fixed fleet size, budgets and routing structures are the most salient. Another, the regulation process, which we have already mentioned, is sequential; first doctors, then transportation coordinators, then regulators, and then finally medical and flight planners decide if, where and how the patient should be moved. (The most important planner, the patient, is not involved after the first step.) By sequential we mean that once the decision at a particular step is made, subsequent steps do not challenge it. The problem is one of world views that do not confront each other. The concept is somewhat alien to military organization, particularly when one organization is subordinate to another. The important question is not "Should we?", but "Can we afford not to?".

REFERENCES

1. A. R. ABDEL-KHALIK AND E. J. LUSK, "Transfer Pricing-A Synthesis", **The Accounting Review** 49, 8-23 (1974).

2. N. I. AGIN AND D. E. CULLEN, "An Algorithm for Transport Routing and Vehicle Loading", in the North-Holland/TIMS Studies in the Management Sciences, Volume 1, **Logistics**, North-Holland Publishing Co., 1975.

3. A. I. ALI AND J. L. KENNINGTON, "The M-Traveling Salesman Problem: A Duality Based Branch and Bound Algorithm", Tech. Report OR 80018, Department of Operations Research, Southern Methodist University, October 1979.

4. __________, R. V. HELGASON AND J. L. KENNINGTON, "An Air Force Logistics Support System Using Multicommodity Network Models", Tech. Report OR 79012, Department of Operations Research, Southern Methodist University, revised February 1981.

5. AMERICAN ACCOUNTING ASSOCIATION, "Information Issues in Managerial Accounting", **The Accounting Review** 47, 317-335 (1972).

6. D. ATKINS, "Managerial Decentralization and Decomposition in Mathematical Programming", **OR Quarterly** 25(4), 615-624 (1974).

7. E. BALAS AND N. CHRISTOFIDES, "A Restricted Lagrangean Approach to the Traveling Salesman Problem", **Mathematical Programming** 21, 19-46 (1981).

8. ________,"An Infeasibility-Pricing Decomposition Method for Linear Programs", **Operations Research** 14(5), 847-873 (1966).

9. L. BARACHET",Graphic Solution of the Traveling Salesman Problem", **Operations Reseach** 5(6), 841-845 (1957).

10. M. S. BAZARAA AND J. J. JARVIS, **Linear Programming and Network Flows**, John Wiley and Sons, New York, 1977.

11. A. BASILEVSKY, **Applied Matrix Algebra in the Statistical Sciences**, North-Holland, New York, 1983.

12. W. J. BAUMOL AND T. FABIAN, "Decomposition Pricing for Decentralization and External Economies", **Management Science** 11(1), 1-32 (1964).

13. M. H. BECKER, The Health Model and Personal Health Behavior, cited in [STCL82], 1974.

14. M. BELLMORE AND S. HONG, "Transformation of the Multisalesmen Problem to the Standard Traveling Salesman Problem," **J. Assoc. Comput. Machinery** 21, 500-504 (1974).

15. ___________ AND J. C. MALONE, "Pathology of Traveling Salesman Subtour Elimination Algorithms", **Operations Research** 19, 278-307 (1971).

16. ___________ AND G. NEMHAUSER, "The Traveling Salesman Problem: A Survey", **Operations Research** 16, 538-558 (1968).

17. L. D. BODIN AND B. L. GOLDEN, "Classification in Vehicle Routing and Scheduling", **Networks** 11, 97-108 (1981).

18. ____________________________, A. ASSAD, AND M. BALL, "Routing and Scheduling of Vehicles and Crews: The State of the Art", **Computers and Operations Research** 10, 62-211 (1983).

19. K. BORCH, "Specification of Objectives in Decision Problems", **Theory and Decision** 1, 5-21 (1970).

20. R. M. BURTON, W. W. DAMON AND D. W. LOUGHRIDGE, "The Economics of Decomposition: Resource Allocation vs. Transfer Pricing", **Decision Sciences** 5, 297-310 (1974).

21. _____________ AND B. OBEL, "The Multilevel Approach to Organizational Issues of the Firm - A Critical Review", **OMEGA** 5(4), 395-414 (1977).

22. G. CARPENTO AND P. TOTH, "Algorithm 548: Solution of the Assignment Problem", **ACM Transactions on Mathematical Software** 6(1), 104-111 (1980).

23. _______________________, "Some New Branching and Bounding Criteria for the Asymmetric Traveling Salesman Problem", **Management Science** 26, 736-743 (1980).

24. R. G. CASSIDAY, M. J. L. KIRBY AND W. M. RAIKE, "Efficient Distribution of Resources Through Three Levels of Government", **Management Science** 17(8), B462-B473 (1971).

25. A. CHARNES AND W. W. COOPER, **Management Models and Industrial Applications**, John Wiley and Sons, Inc., New York, 1961.

26. __________, R. W. CLOWER AND K. O. KORTANEK, "Effective Contrcl through Coherent Decentralization with Preemptive Goals", **Econometrica** 25(2), 1967.

27. N. CHRISTOFIDES, **Graph Theory, An Algorithmic Approach**, Academic Press, New York, 1975.

28. ______________, A MINGOZZI AND P. TOTH, "Exact Algorithms for the Vehicle Routing Problem Based on Spanning Tree and Shortest Path Relaxations", **Mathematical Programming** 20, 255-282(1981).

29. ______________, "The Traveling Salesman Problem", in N. Christofides, P. Toth and C. Sandi, **Combinatorial Optimization**, John Wiley and Sons, Inc., New York, 131-150, 1979.

30. C. W. CHURCHMAN, **The Systems Approach**, Basic Books, New York, 1969.

31. ______________, **The Design of Inquiring Systems**, Basic Books, New York, 1971.

32. ______________, L. AUERBACH AND S. SADAN, **Thinking for Decisions: Deductive Quantitative Methods**, Scientific Research Associates, Inc., Chicago, 1975.

33. ______________, "Philosophical Speculations on Systems Design", **Handbook of Operations Research: Foundations and Fundamentals**, Vol. I, J. J. Moder and S. E. Elmaghraby, Van Nostrand Reinhold Co., New York, 1977.

34. ______________, **The Systems Approach and its Enemies**, Basic Books, New York, 1979.

35. G. CLARKE AND J. WRIGHT, "Scheduling of Vehicles From a Central Depot to a Number of Delivery Points," **Operations Research** 11, 568-581 (1964).

36. J. P. CRECINE, "Defense Budgeting: Organizational Adaptation to Environmental Constraints", in **Studies in Budgeting**, R. Byrne, A. Charnes, W. W. Cooper, C. Davis, and D. Gilford, Eds., North-Holland, Amsterdam, 1971.

37. R. M. CYERT AND J. G. MARCH, **The Behavioral Theory of the Firm**, Prentice-Hall, Inc., Englewood Cliffs, NJ, 1963.

38. F. H. CULLEN, J. J. JARVIS AND H. D. RATLIFF, "Set Partitioning Based Heuristics for Interactive Routing, **Networks** 11, 125-143 (1981).

39. G. DANTZIG AND J. RAMSER, "The Truck Dispatching Problem", **Management Science** 6, 81-91 (1959).

40. __________ AND P. WOLFE, "The Decomposition Algorithm for Linear Programs", **Econometrica** 29(4), 767-778 (1961).

41. __________, **Linear Programming and Extensions**, Princeton University Press, Princeton, N.J., 1963.

42. W. J. DAVIS, "Three-Level Echelon Models for Organizational Coordination", unpublished PhD Thesis, Purdue University, May 1975.

43. Department of Defense. Office of the Assistant Secretary of Defense for Health Affairs. "Military Health Service System: Non-user and User Perceptions and Evaluations", June 1977.

44. ______________________. Department of the Air Force. 375th Aeromedical Airlift Wing. **Mission Briefing.** August 29, 1978.

45. ______________________. Defense Audit Service. **Report on the Audit of the Worldwide Aeromedical Evacuation System**, Case number 8ST-033, 1978.

46. ______________________. **Defense Resource Management Study**, Final Report to the Secretary of Defense, D. D. Rice, US Government Printing Office, Washington, DC, February 1979.

47. ______________________. **Annual Report for Fiscal Year 1982**, Harold Brown, Secretary of Defense, US Government Printing Office, Washington, 1981.

48. ______________________. Department of the Air Force. Air Force Accounting and Finance Center. Personal Statement of Military Compensation, November 1, 1982.

49. ______________________. Joint Chiefs of Staff and the US European Command Cooperating. **Regional Medical Planning Conference After-Action Report**, JCS/J4MM-2-83, January 5, 1983.

50. ______________________. Department of the Air Force. Military Airlift Command. **CONUS Aeromedical Evacuation Study**, Report submitted to the Command Surgeon, March 31, 1983.

51. ______________________. Office of the Assistant Secretary of Defense for Health Affairs. **Defense Health Agency Study.** US Government Printing Office, April 22, 1983.

52. ______________________. The Feasibility and Benefits to be Gained from Creating the Defense Health Agency, Report to the Secretary of Defense, Systems Research and Applications Corp., August 26,1983.

53. ______________________. Department of the Air Force. Military Airlift Command. Command Data Book. April 1984, p.86.

54. T. DOYLE, M. MESTROVICH, J. BECKER, K. CURRY, AND R. PAGE, "Estimating the Populations Served by DOD Hospitals", Paper presented at the ORSA/TIMS Joint National Meeting, Detroit, April 20, 1982.

55. W. L. EASTMAN, "Linear Programming with Pattern Constraints", Unpublished PhD Thesis, Harvard University, July, 1958.

56. A. EILON AND R. FLAVELL, "Note on Many-Sided Shadow Prices", **OMEGA** 2(6), 821-823 (1974).

57. M. M. FLOOD, "The Traveling Salesman Problem", **Operations Research** 4, 61-75 (1956).

58. D. S. FRAGER, J. S. GENUARDI, C. J. SCHUMAKER, JR., M. E. TURNER, AND C. W. WRIGHTSON, JR., "Analysis of Costs at Military Medical Treatment Facilities", Paper presented at the ORSA/TIMS Joint National Meeting, Detroit, MI, April 20, 1982.

59. J. R. FREELAND, "Conceptual Models of the Resource Allocation Decision Process in Hierarchical Decentralized Organizations", PhD Thesis, Georgia Institute of Technology, 215 pages, 1973.

60. _______________, "A Note on a Resource Directive Algorithm for Allocation of Resources in a Decentralized Organization", **Decision Sciences** 6, 186-189 (1975).

61. _______________ AND J. H. MOORE, "Some Organizational Design Implications of Resource Allocation Mechanisms", Research Paper No. 289, Graduate School of Business, Stanford University, June 1977.

62. R. S. GARFINKEL, "On Partitioning the Feasible Set in a Branch-and-Bound Algorithm for the Symmetric Traveling Salesman Problem", **Operations Research** 21(1), 340-343 (1973).

63. W. L. GARRISON AND D. F. MARBLE, "The Structure of Transportation Networks", The Transportation Center, Northwestern University, TCREC Technical Report 62-11, 100 pp. (May 1962).

64. B. GAVISH, "A Note on the Formulation of the M-Salesman Traveling Salesman Problem," **Management Science** 22, 704-705 (1976).

65. __________ AND S. C. GRAVES, "Scheduling and Routing in Transportation and Distribution Systems: Formulations and New Relaxations", Working paper QM8202, Grad. School of Mgmt., Univ. of Rochester, revised August 1982.

66. __________ AND K. SRIKANTH, "An Optimal Solution Method for the Multiple Traveling Salesman Problem," Working Paper QM8027, Grad. School of Mgmt., Univ. of Rochester, 1980.

67. ___________________________, "Efficient Branch and Bound Code for Solving Large Scale Traveling Salesman Problems to Optimality", Working Paper QM8329, Grad. Sch. of Mgmt., Univ. of Rochester, 1983.

68. ___________________________, "Mathematical Formulations for the Dial-A-Ride Problem," Working Paper QM7909, Grad. Sch. of Mgmt., Univ. of Rochester, March 1979.

69. A. M. GEOFFRION, "Elements of Large-Scale Mathematical Programming. Parts I and II.", **Management Science** 16(11), 112-122 (1973).

70. B. E. GILLETT, **Introduction to Operations Research: A Computer-Oriented Algorithmic Approach**, McGraw-Hill, Inc., New York, 1976.

71. ______________ AND J. JOHNSON, "Multi-Terminal Vehicle-Dispatch Algorithm", **OMEGA** 4, 711-718(1976).

72. ______________ AND L. MILLER, "A Heuristic Algorithm for the Vehicle Dispatch Problem," **Operations Research** 22, 340-349 (1974).

73. F. GLOVER, D. KARNEY, D. KLINGMAN AND A. NAPIER, "A Computational Study on Start Procedures, Basis Change Criteria, and Solutions Algorithms for Transportation Problems", **Management Science** 20(5), 793-813 (1974).

74. B. L. GOLDEN, T. L. MAGNANTI, AND H. Q. NGUYEN, "Implementing Vehicle Routing Algorithms," **Networks** 7, 113-148 (1977).

75. _____________, L. BODIN, T. DOYLE, AND W. STEWART, JR, "Approximate Traveling Salesman Algorithms", **Operations Research** 28(3), 694-711 (1980).

76. R. L. GRAHAM, "The Combinatorial Mathematics of Scheduling", **Scientific American** 238, 124-132 (1979).

77. J. E. HAAS, "Transfer Pricing in a Decentralized Firm: A Decomposition Algorithm for Quadratic Programming", **Management Science** 14, B310-B331 (1968).

78. B. HEDBERG AND S. JONSSON, "Designing Semi-Confusing Information Systems for Organizations in Changing Environments", Presented at the Annual Conference of the American Institute for the Decision Sciences, November 10, 1976, 29 pp.

79. M. HELD AND R. KARP, "The Traveling Salesman Problem and Minimum Spanning Trees", **Operations Research** 18, 1138-1162 (1970).

80. "Helicopter Ambulances Too Expensive", News-Democrat (Belleville, IL), June 10, 1983, p. 5.

81. R. HESSE AND R. G. WOOLSEY, **Applied Management Science: A Quick and Dirty Approach**, Scientific Research Associates, Inc., Chicago, 1980.

82. F. S. HILLIER AND G. J. LIEBERMAN, **Introduction to Operations Research**, 3rd. ed., Holden-Day, San Francisco, 1980.

83. J. HIRSCHLEIFER, "Economics of the Divisionalized Firm", **Journal of Business** 30, (1957).

84. S. HONG AND M. PADBERG, "A Note on the Symmetric Multiple Traveling Salesman Problem with Fixed Charges", **Operations Research** 25(5), 871-874 (1977).

85. E. HOROWITZ AND S. SAHNI, **Fundamentals of Computer Algorithms**, Computer Science Press, Rockville, MD, 1978.

86. **The International Mathematical and Statistical Library**, Vol. 3, rev. 9th ed., IMSL, Inc., Houston, 1982.

87. J. JAW, A. R. ODONI, H. N. PSARAFTIS, AND N. H. M. WILSON, "A Heuristic for the Multi-Vehicle Many-to-Many Advance-Request Dial-A-Ride Problem, Working Paper MIT-UMTA-82-3, Mass. Institute of Technology, June 1982.

88. L. P. JENNERGREN, "Studies in the Mathematical Theory of Decentralized Resource Allocation", PhD Thesis, Graduate School of Business, Stanford University, 1971.

89. ________________, "Decentralization on the Basis of Price Schedules in Linear Decomposable Resource Allocation Programs", **Journal of Financial and Quantitative Analysis**, 1407-1415 (1972).

90. ________________, "A Price Schedules Decomposition Algorithm for Linear Programming Problems", **Econometrica** 41, 765-780 (1973).

91. A. JOHNSON, JR., J. T. COOPER AND F. E. ELLEGOOD, "Five-Year Study of Emergency Aeromedical Evacuation in the United States", **Aviation, Space, and Environmental Medicine** 47(6), 662-666 (June 1976).

92. ________________, "Two Years of Routine Patient Movement in the U.S. (Jan. 1974-Dec 1975)", **Aviation, Space, and Environmental Medicine** 48, 451-453 (1977).

93. I. KANT, **Critique of Pure Reason**, 1788.

94. J. L. KENNINGTON, "A Survey of Linear Cost Multicommodity Network Flows", **Operations Research** 26(2), 209-236 (1978).

95. J. KORNAI AND T. LIPTAK, "Two-Level Planning", **Econc metrica** 33(1), 141-169 (1965).

96. __________, **Mathematical Planning of Structural Decisions**, North-Holland Pub. Co., Inc., Amsterdam, 1967.

97. J. S. H. KORNBLUTH, "Accounting in Multiple Objective Linear Programming", **Accounting Review** 59(2), 284-295 (1974).

98. J. B. KRUSKAL, "On the Shortest Spanning Subtree of a Graph and the Traveling Salesman Problem", **Proceedings of the American Mathematical Society** 7, 48-50 (1956).

99. R. E. KUENNE AND R. M. SOLAND, "The Multisource Weber Problem: Exact Solutions by Branch and Bound," **IDA Economic Papers**, H. Williams (ed.), Inst. for Defense Analysis, Arlington VA, 1971.

100. A. W. KUHN, "The Hungarian Method for the Assignment Problem", **Naval Research Logistics Quarterly** 2, 83-97 (1955).

101. G. LAPORTE AND Y. NORBERT, "A Cutting Planes Algorithm for the M-Salesman Problem", **J. Ops. Res. Soc.** 31, 1017-1023 (1980).

102. L. S. LASDON, **Optimization Theory for Large-Scale Systems**, MacMillan and Co., New York, 1970.

103. W. N. LEDBETTER AND J. F. COX, "Are OR Techniques Being Used?", **Industrial Engineering**, 19-21 (1977).

104. L. LEVY, B. L. GOLDEN, AND A. ASSAD, "The Fleet Size and Mix Vehicle Routing Problem", Working Paper MS/S 80-011, College of Business and Management, University of Maryland, College Park, MD, October, 1980.

105. S. LIN AND B. KERNIGHAN, "An Effective Heuristic Algorithm for the Travelling Salesman Problem", **Operations Research** 21, 498-516 (1973).

106. C. E. LINDBLOOM, "The Science of Muddling Through", **Public Administration Review** 19, 79-88 (1959).

107. J. D. C. LITTLE, K. G. MURTY, D. W. SWEENEY, AND C. KAREL, "An Algorithm for the Traveling Salesman Problem", **Operations Research** 11(6), 972-989 (1963).

108. F. C. J. LOKIN, "Procedures for Traveling Salesman Problems with Additional Constraints", **European Journal of Operations Research** 3, 135-141 (1978).

109. J. C. LOWE AND S MORYADAS, **The Geography of Movement**, Houghton Mifflin Co., Boston, 1975.

110. M. MADDOCKS, "Learning Slowly of War's Effects", News-Democrat (Belleville, IL), Jan. 28, 1983, p. 4A.

111. T. L. MAGNANTI AND R. T. WONG, "Network Design and Transportation Planning: Models and Algorithms", **Transportation Science** 18(1), 1-55 (1984).

112. E. MALINVAUD, "Decentralized Procedures for Planning", in **Activity Analysis in the Theory of Growth and Planning**, E. Malinvaud and M. O. L. Bacharach, eds., MacMillan and Co., London, 1967.

113. R. O. MASON, "Basic Concepts for Designing Management Information Systems", **Information for Decision-Making: Quantitative and Behavioral Dimensions**, A. Rappaport, Ed., Prentice-Hall, Inc., Englewood Cliffs, N. J., 2-16, 1975.

114. W. T. McCORMICK, P. J. SCHWEITZER, AND T. W. WHITE, "Problem Decomposition and Data Reorganization in a Clustering Technique", **Operations Research** 20(5), 993-1009 (1972).

115. "McDonnell Gamble Reaps Big Payoff", **St. Louis Post-Dispatch**, 7C, March 1, 1984.

116. R. A. MCFARLAND, **Human Factors in Air Transportation**, McGraw-Hill, Inc., 1953.

117. L. F. McGINNIS, "Implementation and Testing of a Primal-Dual Algorithm for the Assignment Problem", **Operations Research** 31(2), 277-291 (1983).

118. W. MENDENHALL, **Introduction to Linear Models and the Design and Analysis of Experiments**, Wadsworth Pub. Co., Inc., Belmont, CA, 1968.

119. C. MILLER, A. TUCKER, AND R. ZEMLIN, "Integer Programming Formulation of Traveling Salesman Problems," **J. Assn. Comp. Mach.** 7, 326-329 (1960).

120. H. MINTZBERG, "A Note on that Dirty Word "Efficiency"", **Interfaces** 12(5), 101-105 (1982).

121. L. G. MITTEN, "Branch-and-Bound Methods: General Formulation and Properties", **Operations Research** 18, 24-34 (1970).

122. F. W. MOLINA, "A Resource Directive Decomposition Method for Mathematical Programming Based on Lagrangean Theory", PhD Thesis, University of California at Los Angeles, 1977.

123. J. H. MOORE, "Effects of Alternate Information Structures in a Decomposed Organization: A Laboratory Experiment", unpublished research report, Stanford University, November 1976.

124. K. G. MURTY, **Linear and Combinatorial Programming**, John Wiley And Sons, New York, 1976.

125. C. ORLOFF, "Routing a Fleet of M Vehicles to/from a Central Facility," **Networks** 4, 147-162 (1974).

126. C. H. PAPADIMITRIOU AND K. STEIGLITZ, **Combinatorial Optimization: Algorithms and Complexity**, Prentice-Hall, Inc., Englewood Cliffs, N. J., 1982.

127. J. PERL, "A Unified Warehouse Location-Routing Analysis", PhD Thesis, Northwestern University, June 1983.

128. J. PFEFFER AND G. R. SALANCIK, **The External Control of Organizations: A Resource Dependence Approach**, Harper and Row, New York, 1978.

129. R. B. POTTS AND R. M. OLIVER, **Flows in Transportation Networks**, Academic Press, 1972.

130. R. C. PRIM, "Shortest Connection Networks and Some Generalizations", **Bell System Technical Journal** 36, 1389-1401 (1957).

131. H. N. PSARAFTIS, "A Dynamic Programming Approach to the Dial-A-Ride Problem", Report R78-34, Department of Civil Engineering, Massachusetts Institute of Technology, September, 1978.

132. _______________, "A Dynamic Programming Approach to the Single Vehicle Many-to-Many Immediate Request Dial-A-Ride Problem", **Transportation Science** 14(2), 130-154, (1980).

133. _______________, "Two Heuristic Algorithms for the Single Vehicle Many-to-Many Dial-A-Ride Problem", working paper OE-UMTA-81-1, Massachusetts Institute of Techno-logy, October 1981.

134. _______________, "Analysis of an $O(N^2)$ Heuristic for the Single Vehicle Many-to-Many Euclidean Dial-A-Ride Problem", Working Paper, Massachusetts Institute of Technology, January 1981, revised April, 1982. Forthcoming in **Transportation Research.**

135. _______________, "An Exact Algorithm for the Single Vehicle Many-to-Many Dial-A-Ride Problem with Time Windows", Technical Note OE-UMTA-82-2, Massachusetts Institute of Technology, May, 1982.

136. J. B. ROSEN, "Primal Partition Programming for Block Diagonal Matrices", **Numerische Mathematik** 6, 250-260 (1967).

137. T. W. RUEFLI, "Planning in Decentralized Organizations", PhD Thesis, Carnegie-Mellon University, Pittsburgh, 1969.

138. ____________, "Decentralized Transshipment Networks", **Operations Research** 19, 1619-1631 (1971).

139. ____________, "Behavioral Externalities in Decentralized Organizations", **Management Science** 9(5), 649-657, (1971).

140. ____________, "PPBS - An Analytical Approach", in R.F Byrne, A. Charnes, W. W. Cooper, C. A. Davis and D. Gilford, eds., **Studies in Budgeting**, North-Holland Pub. Co., Amsterdam, 1971.

141. _____________, "Analytical Models of Resource Allocation in Hierarchical Multilevel Systems", **Socio-Economic Planning Science** 8, 353-363.

142. R. A. RUSSELL, "An Effective Heuristic for the M-Tour Traveling Salesman Problem With Some Side Constraints," **Operations Research** 25, 517-524 (1977).

143. _____________ AND W. IGO, "An Assignment Routing Problem", **Networks** 9, 1-17 (1979).

144. M. J. SCHIER, "2 Hospitals Plan Aerial Transport", The Houston Post, Oct. 14, 1981, p. 21c.

145. D. A. SCHON, **Technology and Change: The New Heraclitus**, Pergammon Press, New York, N.Y., 1967.

146. T. R. SEXTON, "The Single Vehicle Many to Many Routing and Scheduling Problem", unpublished PhD dissertation, State University of New York at Stony Brook, 1979.

147. D. M. SHAPIRO, "Algorithms for the Solution of the Optimal Cost and Bottleneck Traveling Salesman Problems", unpublished Sc. D. Thesis, Washington University, St. Louis, 1966.

148. H. A. SIMON, **The New Science of Management Decision,** Rev. Ed., Prentice-Hall, Inc., Englewood Cliffs, NJ, 1977.

149. S. J. SIVERD, "A Method for Examining the Impact of Regional Regulation on Hospital Transfers", Paper presented to The Institute of Management Sciences/ Operations Research Society of America Joint National Meeting, Los Angeles, November 13-15, 1978.

150. T. H. C. SMITH, V. SRINIVISAN AND G. L. THOMPSON, "Computational Performance of Three Subtour Elimination Algorithms for Solving Asymmetric Traveling Salesman Problems", **Annals of Discrete Mathematics** 1, 495-506 (1977).

151. V. SRINIVASAN AND G. L. THOMPSON, "Solving Scheduling Problems by Applying Cost Operators to Asignment Models", **Symposium on the Theory of Scheduling and its Applications,** Springer-Verlag, Berlin, 399-425, 1973.

152. N. J. ST CLAIRE, "Policy Impact Model of Military Health Services Utilization", Paper presented at the ORSA/TIMS Joint National Meeting, Detroit, April 20, 1982.

153. D. M. STEIN, "Scheduling Dial-A-Ride Transportation Systems", **Transportation Science** 12, 232-249 (1978).

154. J. A. SVESTKA AND V. E. HUCKFELDT, "Computational Experience With an M-Salesman Traveling Salesman Algorithm," **Management Science** 19, 790-799 (1973).

155. ______________, "Response to 'A Note on the Formulation of the M-Salesman Traveling Salesman Problem'", **Management Science** 22, 706 (1976).

156. C. SWOVELAND, "Decomposition Algorithms for the Multi-commodity Distribution Problem", Unpublished PhD Thesis, University of California, Los Angeles, 1971.

157. A. TEN KATE, "Decomposition of Linear Programs by Direct Distribution, **Econometrica** 40(5), 883--898 (1970).

158. A. THESEN, **Computer Methods in Operations Research**, Academic Press, New York, 1978.

159. F. A. TILLMAN AND T. M. CAIN, "An Upper Bounding Algorithm for the Single and Multiple Terminal Delivery Problem," **Management Science 18**, 664-682 (1972).

160. E. TURBAN, "A Sample Survey of Operations Research Activities at the Corporate Level", **Operations Research** 20, 708-721 (1972).

161. P. VARAYA, "Trends in the Theory of Decision-Making in Large Systems", **Annals of Economic and Social Measurement** 1(4), 493-500 (1972).

162. D. A. WATNE, "Inferences About the Structural Redesign of a Disaggregated Branch Banking System Having Decomposition Externalities", Working Paper CP-400, Center for Research in Management Science, University of California, Berkeley, May 1977.

163. A. WHINSTON, "Pricing Guides in Decentralized Institutions", PhD Thesis, Carnegie-Melon University, Pittsburgh, 1962.

164. M. WIETZMAN, "Iterative Multilevel Planning with Production Targets", **Econometrica** 38(1), 50-65 (1970).

165. A. WILDAVSKY, **Budgeting: A Comparative Theory of Budgetary Processes**, Little, Brown and Co., Boston, 1975.

166. ____________, "Policy Anlysis is What Information Systems Are Not", Working Paper 53, Graduate School of Public Policy, University of California, Berkeley, May 1976, 23 pages.

167. E. V. W. ZSCHAU, "A Primal Decomposition Algorithm for Linear Programming", Graduate School of Business Working Paper 91, Stanford University, January 1967.

INTERVIEWS

1. James Harrell, Commander, USN, MSC, Director, Armed Services Medical Regulating Office, Scott AFB, IL, April 13, 1984. Discussion of patient regulation; rules and procedures; regulator, physicians, and hospital administrator roles; data capture, processing and distribution; and regulation statistics.

2. Jack Jones, Lieutenant Colonel, USAF, Chief, Aeromedical Evacuation Operations Division, Office of the Command Surgeon, Headquarters Military Airlift Command, Scott Air Force Base, IL, February 24, 1982. Broad discussion of aeromedical operations, governing regulations, policies and operating restrictions; system performance data; and cost and budget data.

3. Richard Poff, Lieutenant Colonel, USAF, Chief, Wing Com-mand Post, 375th Aeromedical Airlift Wing, US Air Force Military Airlift Command, Scott Air Force Base, IL, May 10, 1984. Discussion of informal surveys of patients served, resource shortages, proposed revisions of aircraft routing and scheduling procedures, and initiatives to procure additional aircraft.

4. Mr. Dan Reich, Chief of Flight Operations, Flight for Life, Inc., St. Anthony's Hospital, Denver, CO, December 20, 1982. Overview of hospital-based. helicopter air ambulance services in the United States, and the function and operation of Flight for Life.

5. Charles Woods, Lieutenant Colonel, USAF, Chief, Current Operations Division, 375th Aeromedical Airlift Wing, US Air Force Military Airlift Command, Scott Air Force Base, IL, December 20, 1978 and April 6, 1979. Discussion of the aeromedical mission, system organization and functioning, patient and aircraft routing and scheduling, operating rules and procedures, patient movement data, medical crew procedures, and administrative reporting of system performance.

A Systems Approach to the Aeromedical Aircraft Routing Problem using a Computer-based Model

By

Dennis Robert McLain

B.S. (United States Air Force Academy) 1968
M.S. (University of California, Los Angeles) 1969

DISSERTATION

Submitted in partial satisfaction of the requirements for the degree of

DOCTOR OF PHILOSOPHY

in

Business Administration

in the

GRADUATE DIVISION

OF THE

UNIVERSITY OF CALIFORNIA, BERKELEY

Approved: ..C West Churchman...7/10/84...
Chairman Date
..Arie Segev........7/5/84...
..C. Roger Glassey 8/7/84.....

....

A Systems Approach to the Aeromedical Aircraft Routing Problem using a Computer-based Model

by

Dennis Robert McLain